DIGITAL CLASSICS ONLINE

Band 7 (2021)

ISSN: 2364-7957

##DIGITAL CLASSICS ONLINE##

Aus dem Inhalt

Bd. 7 (2021)

##DIGITAL CLASSICS ONLINE##

Editorial

Charlotte Schubert

Die versteckte Macht der Listen

Das Listenwissen, wie es sich als Bestandteil der Schreibsysteme, als Instrument von Konzeptbildungsprozessen und als Mittel der Organisation verschiedener öffentlicher und privater Verwaltungsprozesse seit dem 3. Jahrtausend v. Chr. zeigt, ist außerhalb seines Anwendungsbereichs häufig ignoriert worden, gleichsam als uninteressante Hilfswissenschaft betrachtet worden – zu unrecht, denn nicht nur die sog. schwarzen Listen, Wahllisten oder die Bestsellerlisten für Romane und Sachbücher zeigen, welche immense politische, gesellschaftliche und ökonomische Bedeutung Listen zukommen kann.

Klassisch zur Bedeutung der Listen ist die Feststellung von Jack Goody in „What's in a List?“:[1] „Die Art, wie Wörter (oder ›Dinge‹) in einer Liste angeordnet werden, ist selbst eine Art der Klassifikation oder der Eingrenzung eines ›semantischen‹ Feldes, denn der Vorgang impliziert, daß einzelne Elemente ein- und andere ausgeschlossen werden“.

Folgt man dieser Definition, heißt das: Listen bedeuten Macht und der Ersteller der Liste übt diese Macht aus. Denn sie entscheiden über Einschluß und Ausschluß, über Sichtbarkeit oder Verschwinden, über Erfolg und Mißerfolg bis hin zur Entscheidung über Leben und Tod (Proskriptionslisten, Schindlers Liste). Für politische Listen (Bürgerlisten, Wahllisten, schwarze Listen etc.) ist in der Regel bekannt, wer die Listen formal aufstellt und zu welchem Zweck sie erstellt werden. Aber wie sieht dies im großen Feld der Digitalisierung im Zeitalter der Digitalität aus? Wer kontrolliert die Listen und ihre Zirkulation, die doch heute zunehmend Grundlage von Datenbanken, Speicher- und Auswertungsprozessen etc. sind?

Es wird derzeit viel über die veränderten medialen Übertragungsverhältnisse, über die veränderten medialen Bedingungen, über Netzpolitik etc. diskutiert: Das Sammeln von Daten, die Berechtigungen von Auswertungen und Kommerzialisierungen, auch institutionelle Überwachungspraktiken spielen im öffentlichen Diskurs eine zunehmend große Rolle. Aber eine Grundlage dieser Praktiken wird weitestgehend ignoriert: die Liste, oder anders ausgedrückt die Liste in der Form von Indizes. Wir haben es in unserer zunehmend digitalisierten Welt mittlerweile mit Listen einer spezifischen Form zu tun: den Indizes, die Datenbanken zugrunde liegen und die die Suche und das Sortieren in den Daten ermöglichen. Diese Indizes haben nicht nur eine Verweis- oder Zeigefunktion, indem Einträge (Zeiger) die Beziehung zwischen den Daten und ihrer Anordnung definieren, so daß bei einer Abfrage der Daten die gesuchten Datensätze anhand dieser Zeiger gefunden werden. Sie sind alles andere als ein passives Werkzeug, das

[1] Goody (2012), 384. Zu poetischen Listen vgl. z.B. Eco (2009), 118.

lediglich zur Auffindung von Informationen dient, sondern, so Stäheli, „Teil eines Prozesses ..., der die Daten strukturiert“, die indiziert werden.[2]

Zugrunde liegt eine normierende Struktur, die über ihre Verweisfunktion Daten erschließen und somit Wissen schaffen soll. Man kann nun unterschiedliche Dimensionen von Listen unterscheiden. Diese Dimensionen sind für den Nutzer häufig nicht sichtbar, wobei aber gerade diese Unsichtbarkeit nach landläufiger Meinung zu ihrer Effizienz, die meist sogar mit handfesten geschäftlichen (z.B. bei Amazons sog. Bestsellerlisten oder bei den kuratierten Playlisten der Streaming-Anbieter) oder politischen Interessen (früher die Zensuslisten, heute z.B. die Listen aus Geheimdienstabfragen) verbunden wird. Der Code, der Algorithmus, der der Erstellung der Liste bzw. des Indexes zugrunde liegt, wird entweder gar nicht offengelegt oder er ist für andere, insbesondere die Nutzer, schlicht unverständlich. In jedem Fall bleibt dieser Prozeß der Erstellung unsichtbar.[3] Der Akt der Erstellung selbst wiederum – man denke an die Ergebnisse einer Suchanfrage im Internet – gibt keinerlei Hinweis auf die ihm inhärenten Prozesse: Das Ergebnis wird präsentiert und geht wie eine evidente Tatsache in die Nutzung ein.

Die Listen, die zur Erstellung einer Auswertung, eines Ergebnisses o.ä. führen, erzeugen ein eigenes Narrativ wie z.B. die Unterstellung von Gefahrensituationen oder im geschäftlichen Bereich die Steuerung des Konsumverhaltens von Nutzern über Rankings. Die mangelnde Transparenz im Hinblick auf Kontrolle, Kontrolleure und Kontrollierte gefährdet das gesamtgesellschaftliche Gefüge unserer digitalisierten Zivilisation, weil die epistemologischen Grundlagen des automatisierten Listings in der digitalen Welt meist verschleiert, verborgen oder doch zumindest ohne qualitative oder politische oder überhaupt gesellschaftliche Kontrolle angewendet werden.

Entscheidend ist immer, welche Wissensordnung reproduziert, reflektiert oder geschaffen wird und dann in eine Liste eingeht. Eine automatisiert erstellte Liste beruht auf statistischen Kalkulationen und Analysen von Mustern (z.B. der Annahme, daß ähnliche Begriffe auf ähnliche Muster führen). Die Wissensordnungen, die diesen Kalkulationen und Analysen zugrunde liegen, sind Ausdruck verschiedenster, möglicherweise nicht immer unberechtigter Interessen – doch ihnen allen ist gemeinsam, daß sie – anders als klassische, aufgeschriebene Listen – mit den herkömmlichen grammatikalischen oder narrativen Zusammenhängen gebrochen haben, da sie automatisch auf der Grundlage von Algorithmen erzeugt werden, deren genaue Intention oder Wirkweise, wie bereits betont, weder bekannt oder nachvollziehbar ist, die jedoch einen beträchtlichen Teil unseres Lebens steuern, überwachen und kontrollieren. In jedem Fall läßt sich festhalten, daß durch diese geheime Macht der Listen die vielgerühmte Öffentlichkeit und Gemeinschaftlichkeit des Internets als Medium[4] komplett unterlaufen werden.

Besonders wirksam ist der von Google automatisch erzeugte Index, aus dem nach dem Google Algorithmus ein PageRank erstellt wird, der bei Suchanfragen die Ergebnisse auflistet.[5] Das Ziel dieses

[2] Stäheli (2019), 37. Die hier formulierten Überlegungen verdanken dem Beitrag von Stäheli (2016 [dt. 2019]) eine Menge: Seine Ausführungen zu der Historie des Indizierens und dem Diskurs über das Indizieren gehören zu den bisher eher seltenen Arbeiten, die die Problematik der „Politik von Listen“ nicht nur im Hinblick auf literaturwissenschaftliche oder soziologische Implikationen untersuchen, sondern auch in Bezug auf die Digitalität analysieren. Das Thema selbst ist auf der Jahrestagung der Mommsen-Gesellschaft 2019 in Berlin von Hannes Kahl und mir unter dem Titel „Liste und Index: Zur Überführung des Analogen ins Digitale in den Klassischen Altertumswissenschaften“ vorgestellt worden und Bestandteil eines DFG-geförderten Forschungsprojektes, zu dem noch weitere Publikationen folgen werden.

[3] Stäheli (2019), 19.

[4] Stalder (2016) in Kapitel 2: Formen der Digitalität.

[5] 1998 veröffentlichten Brin und Page an der Stanford University ein Paper mit dem Titel „The Anatomy of a Large-Scale Hypertextual Web Search Engine“. Hier wurde zum ersten Mal der „PageRank“ erwähnt, die Technologie, mit der Google bis heute Suchergebnissen einen Rang zuordnet: http://infolab.stanford.edu/~backrub/google.html (abgerufen am 22.01.2021).

PageRank haben die Entwickler Page und Brin 1998 bereits öffentlich formuliert: Werbung sei das entscheidende Geschäftsmodell für kommerzielle Suchmaschinen. Gleichwohl würden die Ziele des Werbebranchenmodells nicht immer mit der gewünschten Qualität übereinstimmen.[6]

Wie hat sich nun Google zu diesem Anspruch auf Qualität positioniert? Google läßt jährlich ca. 40.000 sogenannte „Präzisionsauswertungen“ durch „Bewerter“ durchführen, bei denen diese „Suchbewerter“ die Qualität der Ergebnisse für verschiedene Suchen ermitteln.[7] Dazu hat Google auch Richtlinien veröffentlicht, aus denen die Kriterien für die Bewertung von Ergebnissen ersichtlich werden. Dieses 175 Seiten lange Dokument[8] führt z.B. unter der Überschrift „Your Money or Your Life“ (abgekürzt als YMYL zitiert) auf, daß solche Bewertungen Auswirkungen auf Gesundheit und Finanzen von Benutzern haben können.[9] Das ist sicher richtig, nur: Ist es richtig, noch dazu, wenn hier Geld und Leben alternativ gestellt werden, darüber Google entscheiden zu lassen? Der krönende Abschluß dieser Entwicklung war dann die Entscheidung von Google[10] im Frühjahr 2016, daß die Öffentlichkeit nicht mehr in der Lage sein solle, die PageRank-Technologie und deren Daten einzusehen!

Man könnte durchaus zu der Vermutung kommen, daß die digitale Transformation der Gesellschaft zu einer „List Culture“ führen könnte, in der – trotz aller Debatten und Gesetzesinitiativen – die geheime Macht der Listen regiert.

[6] http://infolab.stanford.edu/~backrub/google.html (abgerufen am 22.01.2021).

[7] https://blog.hubspot.de/marketing/google-algorithmus (abgerufen am 22.01.2021).

[8] https://static.googleusercontent.com/media/guidelines.raterhub.com/de//searchqualityevaluatorguidelines.pdf (abgerufen am 22.01.2021).

[9] So heißt es a.a.O.: „Users need high quality information from authoritative sources when researching products, especially when products are expensive or represent a major investment/important life event (e.g., cars, washing machines, computers, wedding gifts, baby products, hurricane shutters, large fitness equipment). When buying products, users need websites they can trust: good reputation, extensive customer service support, etc. Results for product queries may be important for both your money and your life (YMYL)!“

[10] https://searchengineland.com/rip-google-pagerank-retrospective-244286 (abgerufen am 22.01.2021).

Literatur

Eco (2009): U. Eco, Die unendliche Liste, München 2009.

Goody (2012): J. Goody, Woraus besteht eine Liste?, in: S. Zanetti (Hrsg.): Schreiben als Kulturtechnik, Grundlagentexte, Berlin 2012, 338–396.

Stäheli (2019): U. Stäheli, Indizieren – Die Politik der Unsichtbarkeit, in: M. Stempfhuber / E. Wagner (Hrsgg.), Praktiken der Überwachten, Öffentlichkeit und Privatheit im Web 2.0, Wiesbaden 2019, 17–41 (= Indexing – The politics of invisibility, in: Environment and Planning D: Society and Space 31:1, 2016, 14–29).

Stalder (2016): F. Stalder, Kultur der Digitalität, Frankfurt 2016.

Young (2017): L. C. Young, List Cultures: Knowledge and Poetics from Mesopotamia to BuzzFeed, Amsterdam 2017.

Internetseiten

http://infolab.stanford.edu/~backrub/google.html (abgerufen am 22.01.2021).

https://blog.hubspot.de/marketing/google-algorithmus (abgerufen am 22.01.2021).

https://searchengineland.com/rip-google-pagerank-retrospective-244286 (abgerufen am 22.01.2021).

https://static.googleusercontent.com/media/guidelines.raterhub.com/de//searchqualityevaluatorguidelines.pdf (abgerufen am 22.01.2021).

Autorenkontakt[11]

Prof. Dr. Charlotte Schubert

Lehrstuhl für Alte Geschichte
Historisches Seminar
Universität Leipzig
Beethovenstr. 15
04107 Leipzig

Email: schubert@uni-leipzig.de

[11] Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren. Alle Inhalte dieses Beitrages unterstehen, soweit nicht anders gekennzeichnet, der Lizenz CC BY 4.0.

##DIGITAL CLASSICS ONLINE##

Creating the First Digital *Handbook of Latin Phonetics*: Between Linguistics, Digital Humanities and Language Teaching

Tommaso Spinelli

Abstract: This article discusses the creation of an innovative e-learning resource that provides a unique breadth of frequency, grammatical, and phonetic information on both Classical and Ecclesiastical Latin. Designed to bridge teaching and research, this new digital toolkit, which is available as both an online program and an Android mobile app, provides a frequency list of the most common Latin lemmas, as well as phonetic and grammatical information, including their syllabication, accentuation, and Classical and Ecclesiastical phonetic transcription according to the standards of the International Phonetic Alphabet. After providing a concise overview of the different ways in which Latin was and still is pronounced, this article will discuss the methodological and practical issues faced by the creation of the toolkit from the choice of an effective lemmatizing technique for identifying and categorizing inflected word-forms, to the creation of algorithms to accentuate Latin lemmas and transcribe Latin sounds (potentially involving multiple characters of the Latin alphabet) into IPA characters. In so doing, it will offer insights into the technologies used to maximize the impact of this new e-learning resource on teaching and research.

Introduction

This article discusses the recent creation of the first online *Handbook of Latin Phonetics*, an innovative opensource digital toolkit that provides a unique breadth of frequency, grammatical, and phonetic information on both Classical and Ecclesiastical Latin. Originally conceived within the award-winning project *Latine Loquamur* (undertaken to support the reform of classical language teaching at the University of St Andrews), the digital toolkit described in this article was developed at the *Pontificium Institutum Altioris Latinitatis* (Pontifical Salesian University of Rome), to meet the needs of the ever-increasing number of scholars and students who study Latin in Latin, or who focus on late-antique texts.[1] Accordingly, the *Handbook of Latin Phonetics* toolkit currently provides, for the first time ever, a frequency

[1] The *Latine Loquamur Project* was designed by Tommaso Spinelli in collaboration with Alice König, Giuseppe Pezzini and Giacomo Fenzi, and was awarded funding by the Teaching Development Office of the University of St Andrews in 2018. This project involved the creation of an online *Dictionary of Latin Synonyms*, which was published by the University of St Andrews in December 2018 (https://doi.org/10.17630/3cf644e6-86b8-44d0-a50a-b33c7ca86072; last access 17.10.2020), and of other e-learning resources (e.g. Moodle presentations, exercises, and interactive games) for the study of Latin that will be discussed in another article for reasons of space. The *Handbook of Latin Phonetics* presented in this article is available as both an app and a program: the app was developed by Tommaso Spinelli during his Postdoc at the Pontifical Salesian University of Rome in collaboration with Cleto Pavanetto, Giacomo Fenzi, Kamil Kolosowski, and Jan Rybojad, and was published – thanks to the collaboration of Miran Sajovic – by the Pontifical Salesian University of Rome in 2020. (https://play.google.com/store/apps/details?id=com.kolosowski.latinhandbook; https://doi.org/10.17630/19ce37ba-2d35-4920-bd7f-6287977de369; last access 17.10.2020). The online version of the *Handbook of Latin Phonetics*, which was developed by Tommaso Spinelli with the informatic assistance of Giacomo Fenzi at the University of St Andrews, is currently hosted in the GitHub repository of the Latine Loquamur Project (https://github.com/latineloquamur?tab=repositories; last access 17.10.2020) and can be found in the folder titled Latineloquamur-toolkit-IPA-transcriber-and-App. In the same repository users can find also the Dictionary of Latin Synonyms, which is not discussed in this article, and the link to download its app (https://github.com/latineloquamur/dictionary-of-latin-near-synonyms; last access 17.10.2020).

list of the 6500 most common Latin lemmas as attested in the entire extant corpus of Latin literature, as well as unique phonetic and grammatical information, including their syllabication, accentuation, and Classical and Ecclesiastical phonetic transcription according to the standards of the International Phonetic Alphabet.

This toolkit, which is available as both a RUST program (referred to as *Latineloquamur-toolkit-IPA-transcriber-and-App* in GitHub) and an Android mobile app (titled Handbook of Latin Phonetics), faced significant methodological and practical issues during its creation and development, such as the choice of an effective lemmatizing technique for identifying and categorizing inflected word-forms, the creation of algorithms to accentuate Latin lemmas, and the development of an innovative program to transcribe Latin sounds (potentially involving multiple characters of the Latin alphabet) into IPA characters corresponding to different pronunciations of Latin.[2] After providing a concise overview of the different ways in which Latin was and still is pronounced, this article will discuss the complex interaction between linguistics, phonology, and digital humanities. It will explore the methodologies and principal technologies used within this digital project to offer rigorous frequency and phonological information on Latin lemmas, and to maximize its impact on teaching and research.

Pronouncing Latin: between teaching and research

One of the aims of the *Latin Phonetics* digital toolkit is to further the creation of a shared rigorous methodology for the pronunciation of Latin lemmas, and for the identification of the words most used by the Latin authors that a given student or researcher might want to prioritize in their studies. Both 'frequency' and 'pronunciation' have played a key role in language teaching and rhetorical studies since antiquity. Latin authors such as Cicero, Varro, and Quintilian often referred to the *usus* (use) of a word or to its frequency in their literary, grammatical, and stylistic discussions.[3] Similarly, the pseudo-Cicero's *Rhetorica ad Herennium* devotes an entire section to the role of pronunciation in 'delivering' a speech (3.19.1–2), which is also discussed by Quintilian in his *Institutio Oratoria* (1.4; 1.7), while, in the third century CE, the grammarian Probus encourages his students to pronounce correctly the words *speculum* (mirror) and *columna* (column), avoiding the wrong forms *speclum* and *colomna*.[4] And yet, despite the importance of such themes, not enough attention has been paid to them by modern digital scholarship. While the last couple of decades have seen the publication of many frequency dictionaries for modern languages, no comprehensive frequency dictionary yet exists for Latin, as the few modern attempts to provide rigorous lemmatization and counts of Latin words have treated very limited textual corpora, and have adopted remarkably different methodologies, as we shall see better in the following analysis.[5]

Even more problematic is the situation concerning the pronunciation of Latin. Ancient literary and documentary sources indicate that Latin was spoken differently synchronically at different stages of

[2] The two different names of the program and the app are due to the different stages of the development of the toolkit and to the different institutions that published those tools, the University of St Andrews and the Pontifical Salesian University of Rome respectively. However, to avoid confusion in this article I will refer to these tools as the *Latin Phonetics* app/program.

[3] Joseph Denooz (2010), 1–2 has shown that the word *usus* ('use') is used to explain linguistic facts 45 times by Varro in his *De lingua Latina*, 163 times by Cicero in the *De Oratore* and the *Orator*, and 163 times in Quintilian's *Institutio Oratoria*. Moreover, Quintilian uses the adjective *frequens* ('frequent') and the adverb *frequenter* ('frequently') some 223 times in his linguistic and stylistic considerations. Cf. also Cic. *De Inv*. 1.9.4; 1.9.10; *De Or*. 3.140.4.

[4] The so-called list of the '*appendix Probi*' has been variously dated to the third or the fifth century CE. See Barnett (2006), 257–278.

[5] See, for instance Diederich (1939), Delatte/Evrard/Govaerts/Denooz (1981), Denooz (2010).

Roman history and in different regions of the empire by different social classes. For instance, Lucilius jokes about the rustic pronunciation of a certain *C**ae**cilius*, who was *pr**ae**tor urbanus* (urban pretor), by saying, in a phonetic spelling, ‘*C**e**cilius pr**e**tor ne rusticus fiat*’ (Let Cecilius not be a rustic pretor; Lucil. 1130, M.), remarking on the fact that, as we know from Varro (L. 5.97), the diphthong ae was already pronounced /e/ in the countryside in Classical times.[6] Epigraphs show the existence of different pronunciations of Latin throughout the history of Rome, and the *Historia Augusta* (*Hadr*. 3.1) recounts that the emperor Hadrian (117–38 CE) was mocked for his Hispanic accent.[7] This ancient diversity has been only partially reduced in modern times; it has therefore been an urgent and challenging necessity to create a tool able to provide a standardized pronunciation of Latin.[8]

Although the first *Congrès International Pour le Latin Vivant* (the first international conference for living Latin), held in Avignon in 1956, tried to foster a shared Classical pronunciation of Latin in modern times, at least three different ways of reading Latin are still commonly – and often unthinkingly – used by different institutions.[9] The first way is the so-called ‘national’ because of its proximity to the phonetic system of the modern languages of the countries in which Latin is read.[10] According to this pronunciation, for example, the lemma *Caesar*, which was pronounced /’kaɛ̯.sar/ in Classical Latin and /’tʃɛ.sar/ in Ecclesiastical, is read as /’tʃɛ.sar/ in Italy, /ʃɛ.’sar/ in France, /’sɪ.sar/ in Britain, and /’tsɛ.sar/ in Germany. The second way is the so-called ‘Ecclesiastical’ because it is officially used by the Catholic Church. Although it looks similar to the Italian pronunciation of Latin, this pronunciation is supranational and reflects the diction of Latin used in Rome during the fourth and fifth centuries CE.[11] The third way is the so-called ‘Classical’ pronunciation or ‘*restituta*’. Starting from the Renaissance period, this system of pronunciation used the phonetic clues provided by ancient grammatical texts and epigraphs to reconstruct the language arguably spoken by cultured Romans in the first century BCE and the first century CE.[12] A further complicating factor is that, while an ever-increasing number of institutions worldwide has started to teach Latin in Latin, using the Ørberg’s and Cambridge’s textbooks that encourage a more active use of the language in its ‘Classical’ pronunciation, other world-leading institutions (such as the Salesian University of Rome and the *Pontificium Institutum Altioris Latinitatis*) have continued to use the Ecclesiastical pronunciation that is also used to read late-antique and early-medieval texts, to which Classicists have increasingly shifted their attention in the last two decades.[13]

At this critical juncture, my new toolkit builds upon recent developments in the fields of digital humanities and Latin linguistics to provide students worldwide with a rigorous guide to the pronunciation of both Classical and Ecclesiastical Latin. In particular, while the Latin dictionaries currently available in many countries tend to provide only the quantity (or length) of the penultimate syllable of lemmas, the *Latin Phonetics* program and app provide more complete information on the accentuation, prosody, syllabication, and IPA phonetic transcription of Latin lemmas. The following analysis will explore the

[6] See Ramage (1963), 390–414.

[7] See, for example, the commonly attested form coss. for *consules*, or the names *Crescentsianus* and *Vincentza* respectively attested in *CIL* XIV, 246; VII, 216. On dialectal pronunciations of Latin see Oniga (2003), 39–62.

[8] On the Church’s use of Latin see the epistle *Romani Sermonis* by Paulus VI (1976). On the use of Latin in modern academia see Short/George (2013). On the bidirectional influence of national languages on Latin, see Serianni (1998), 27–45.

[9] See Allen (1966), 102; Pavanetto (2009), 9–10; Traina/Bernardi-Perini (1998), 22–29.

[10] See Collins (2012).

[11] An overview of the most important features of Ecclesiastical Latin is provided by Collins (1988).

[12] See, for instance, Erasmus’ *De recta Latini Graecique sermonis pronuntiatione* (1528). On this theme see also Allen (1966); Oniga (2014).

[13] See the overview provided by Chiesa (2012) and Spinazzé (2014), but also the seminal work on stylometry of the ‘Quantitative Criticism Lab’ (https://www.qcrit.org/researchdetail/kHXib8DissfMp53Yx; last access 17.10.2020). See also Harrington/Pucci (1997), Avitus (2018), and Norberg (1999).

methodology and innovative technologies used to create this tool, discussing the potential and the limits of the digital technologies that can be currently deployed to process the Latin language.

Existing technologies

The *Latin Phonetics* toolkit builds on and bridges together several different technologies developed in recent years to meet the new need for more complete phonetical information which can be used not only to speak and write in Latin, but also to study rhythmic prose texts and the style of authors. Leaving aside the work-in-progress Latin dictionary on Wikipedia that unsystematically offers some information on the pronunciation of Latin words, the best-equipped tool currently available is that offered by the Classical Language Toolkit Project (CLTK).[14] This international opensource project offers both a 'macronizer' and a 'phonetic transcriber'. Based on an original algorithm developed by Johan Winge in 2015, the macronizer can mark Latin vowels according to their length, using a POS tagger which matches words with the lexical entries of Morpheus.[15] Although this tool does not provide accentuation of lemmas and has an accuracy of around 86.3% (depending on which of the three available POS is used), it allows a more complete prosodic mark-up than that usually offered by traditional dictionaries.[16] Moreover, the 'phonetic transcriber' represents the first attempt to provide a rigorous phonetic transcription of the Latin language according to the IPA standards.[17] This tool transliterates Latin lemmas into their phonetic forms using a list of replacements based on Allen's reconstruction of the phonetics of Classical Latin (1966). However, while the source codes of both these tools are available on Github, they do not offer a user-friendly interface, so that only expert users, with a good knowledge of programming, can actually use them to process Latin words. Moreover, the CLTK phonetic transcriber only provides information on the Classical pronunciation of Latin. Similarly, the project *LatinWordnet2.0*, which is being developed at the University of Exeter by William Short, provides the Classical phonetic transcription for Latin lemmas, but this data is currently accessible only to expert users.[18] Although they do not provide a phonetic transcription of Latin, it is worth mentioning other programs which have tried to address similar issues. The first is *Google Translate*, which now offers the Ecclesiastical pronunciation (but not accentuation and IPA transcription) of Latin words.[19] The second is *Collatinus*, which was developed within the project *Biblissima* for the study of medieval and modern texts, and can divide by syllable and accentuate Latin lemmas or small texts.[20] The third is the *Quantitative Criticism Lab* that, while not providing the pronunciation of Latin lemmas, offers detailed information on the prosody of single words and entire texts, using quantitative metrics to support both linguistic and stylistic analysis.[21] Similar are the projects *Cursus in Clausula*, developed at the University of Udine, which detects the quantitative and tonic

[14] On the dictionary offered by Wikipedia see: https://en.wiktionary.org/wiki/Wiktionary:Main_Page (last access 26.10.2020). The CLTK project is a Python library containing tools for the natural language processing (NLP) of ancient Eurasian languages: http://cltk.org/ (last access 02.09.2020).

[15] The algorithm and its explanation are available at https://cl.lingfil.uu.se/exarb/arch/winge2015.pdf (last access 02.09.2020); Morpheus is a morphological parsing and lemmatizing tool integrated into the Perseus Project http://www.perseus.tufts.edu/hopper/ (last access 02.09.2020).

[16] The Python macronizer is available at https://github.com/cltk/cltk/blob/master/cltk/prosody/latin/macronizer.py (last access 02.09.2020).

[17] The CLTK transcriber can be accessed at https://github.com/cltk/cltk/blob/master/cltk/phonology/latin/transcription.py (last access 02.09.2020).

[18] Cf. https://github.com/wmshort/latinwordnet-archive; https://latinwordnet.exeter.ac.uk/ (last access 02.09.2020).

[19] See https://translate.google.com/?sl=la#view=home&op=translate&sl=la&tl=en&text (last access 02.09.2020).

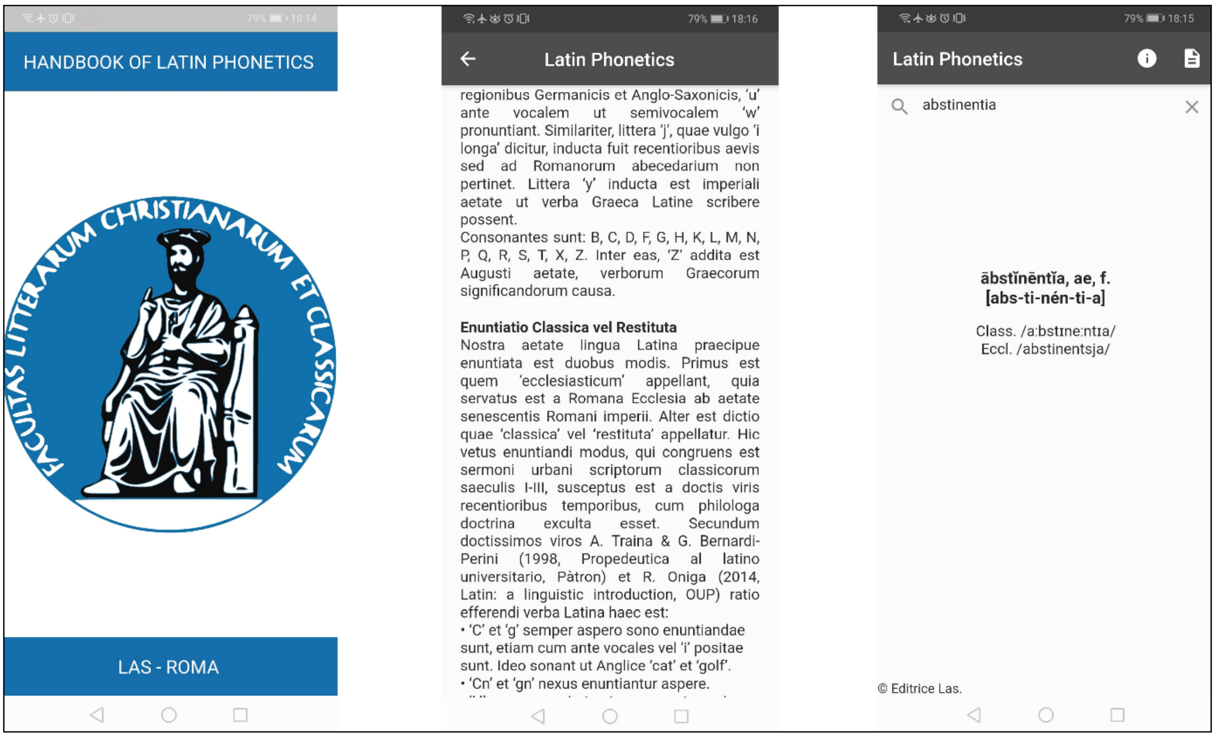
[20] The codes are available at https://github.com/biblissima/collatinus (last access 02.09.2020). See also https://projet.biblissima.fr/ (last access 02.09.2020).

[21] See https://www.qcrit.org/ (last access 02.09.2020).

rhythm of prose *clausulae*, and the toolkit of *Pedecerto* that, developed within the project *FIRB Traditio Partum* by the University Ca' Foscari of Venice, can perform automatic scansion of Latin verses.[22]

The first app of Latin phonetics: outline and features

Distinct from the previous contributions described above, the *Handbook of Latin Phonetics* has been designed to bring together teaching and research. Accordingly, it aims to advance the automated processing of the Latin language through the creation of original algorithms for a rigorous phonetic transcription of both Classical and Ecclesiastical Latin. It also aims to provide students, teachers, and researchers across the world with a compact, freely accessible, and easy-to-use toolkit to study Latin in Latin. For this reason, the *Handbook of Latin Phonetics* has been made available both as an online opensource program for expert users (discussed in detail in the following section) and as a user-friendly Android app (discussed in this section) which, developed in collaboration with Kamil Kolosowski, displays and provides the most important features of the toolkit in an accessible format.



regionibus Germanicis et Anglo-Saxonicis, 'u' ante vocalem ut semivocalem 'w' pronuntiant. Similariter, littera 'j', quae vulgo 'i longa' dicitur, inducta fuit recentioribus aevis sed ad Romanorum abecedarium non pertinet. Littera 'y' inducta est imperiali aetate ut verba Graeca Latine scribere possent.
Consonantes sunt: B, C, D, F, G, H, K, L, M, N, P, Q, R, S, T, X, Z. Inter eas, 'Z' addita est Augusti aetate, verborum Graecorum significandorum causa.

Enuntiatio Classica vel Restituta
Nostra aetate lingua Latina praecipue enuntiata est duobus modis. Primus est quem 'ecclesiasticum' appellant, quia servatus est a Romana Ecclesia ab aetate senescentis Romani imperii. Alter est dictio quae 'classica' vel 'restituta' appellatur. Hic vetus enuntiandi modus, qui congruens est sermoni urbani scriptorum classicorum saeculis I-III, susceptus est a doctis viris recentioribus temporibus, cum philologa doctrina exculta esset. Secundum doctissimos viros A. Traina & G. Bernardi-Perini (1998, Propedeutica al latino universitario, Pàtron) et R. Oniga (2014, Latin: a linguistic introduction, OUP) ratio efferendi verba Latina haec est:

- 'C' et 'g' semper aspero sono enuntiandae sunt, etiam cum ante vocales vel 'i' positae sunt. Ideo sonant ut Anglice 'cat' et 'golf'.
- 'Cn' et 'gn' nexus enuntiantur aspere.

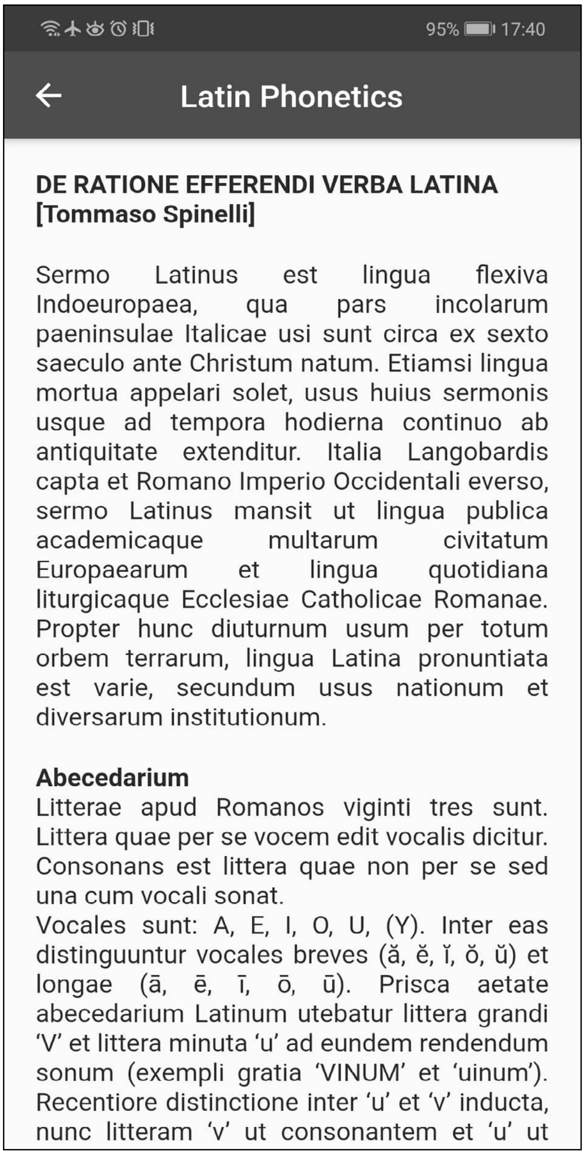
ābstĭnēntĭa, ae, f.
[abs-ti-nén-ti-a]

Class. /aːbstɪneːntɪa/
Eccl. /abstinentsja/

Fig. 1 Latin Phonetics App.

[22] See respectively http://cursusinclausula.uniud.it/public/ (last access 02.09.2020) and http://www.pedecerto.eu/public/ (last access 02.09.2020).

The Android app, freely available on Google Play, is organized as follows: 1) a learning section, containing an introduction to Latin phonetics and a list of the most frequently attested Latin lemmas, 2) a search tool offering a wide range of prosodic and phonetic information on Latin lemmas, and 3) an 'info' section providing details on the app and its related programs.[23]



DE RATIONE EFFERENDI VERBA LATINA [Tommaso Spinelli]

Sermo Latinus est lingua flexiva Indoeuropaea, qua pars incolarum paeninsulae Italicae usi sunt circa ex sexto saeculo ante Christum natum. Etiamsi lingua mortua appelari solet, usus huius sermonis usque ad tempora hodierna continuo ab antiquitate extenditur. Italia Langobardis capta et Romano Imperio Occidentali everso, sermo Latinus mansit ut lingua publica academicaque multarum civitatum Europaearum et lingua quotidiana liturgicaque Ecclesiae Catholicae Romanae. Propter hunc diuturnum usum per totum orbem terrarum, lingua Latina pronuntiata est varie, secundum usus nationum et diversarum institutionum.

Abecedarium
Litterae apud Romanos viginti tres sunt. Littera quae per se vocem edit vocalis dicitur. Consonans est littera quae non per se sed una cum vocali sonat.
Vocales sunt: A, E, I, O, U, (Y). Inter eas distinguuntur vocales breves (ă, ĕ, ĭ, ŏ, ŭ) et longae (ā, ē, ī, ō, ū). Prisca aetate abecedarium Latinum utebatur littera grandi 'V' et littera minuta 'u' ad eundem rendendum sonum (exempli gratia 'VINUM' et 'uinum'). Recentiore distinctione inter 'u' et 'v' inducta, nunc litteram 'v' ut consonantem et 'u' ut

Fig. 2 The app's introductory section.

The first page of the app is an introductory section that, divided in two parts, contains material for the independent e-learning of Latin. The first part, titled *De Ratione Efferendi Verba Latina* offers a brief history of the Latin language, basic notions of linguistics and phonetics, and an up-to-date explanation (written in Latin) of the main differences between Classical and Ecclesiastical pronunciations. This explanation deals especially with the differences in sound between the diphthongs *ae* and *oe* (which are pronounced as monophthongs in Ecclesiastical Latin), and with the pronunciation of velar and voiced plosives (*c*, *g*), which are never soft in Classical Latin, and of the group '-*ti* + vowel', generally pronounced /tɪ/ in Classical Latin and /tsi/ in Ecclesiastical.[24] Different from many modern grammars and digital programs based on Allen's Vox Latina (1966) for Classical Latin, and Nunn's Introduction to Ecclesiastical Latin (1927), this explanation builds on more recent studies such as those on Classical Latin by Traina/Bernardi-Perini (1998) and Oniga (2014), and those on Ecclesiastical Latin by Collins (1998) and especially Pavanetto (2009), who was the head of the Pontifical Institute Latinitas governing the Catholic Church's official use of Latin. This approach governs the phonetic transcription performed by the program, which is summarized in the following table (table 1), especially concerning the sounding of diphthongs in Classical Latin. In this respect, the development of the so-called historical and generative grammar over the past century has revealed that the diphthongs *ae* and *oe* evolved from the older forms *ai* and *oi*, which left a mark on the spelling used on some epigraphs composed before the end of the second century CE. For instance, in the inscription adorning the tomb of Scipio Barbatus, dated around the 250 BCE, the term ***ae**dilis* ('aedile', the censor aedilis was an elected officer responsible for the maintenance of public buildings) is spelled ***ai**dilis* (*CIL* 06, 01287).

[23] The app can be downloaded at https://play.google.com/store/apps/details?id=com.kolosowski.latinhandbook (last access 15.10.2020). The online program is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App (last access 02.09.2020). The constitutive elements of the app can be inspected through this link: https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/Mobile%20APP (last access 02.09.2020).

[24] The introduction is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/Mobile%20APP/Intro_App.txt (last access 02.09.2020).

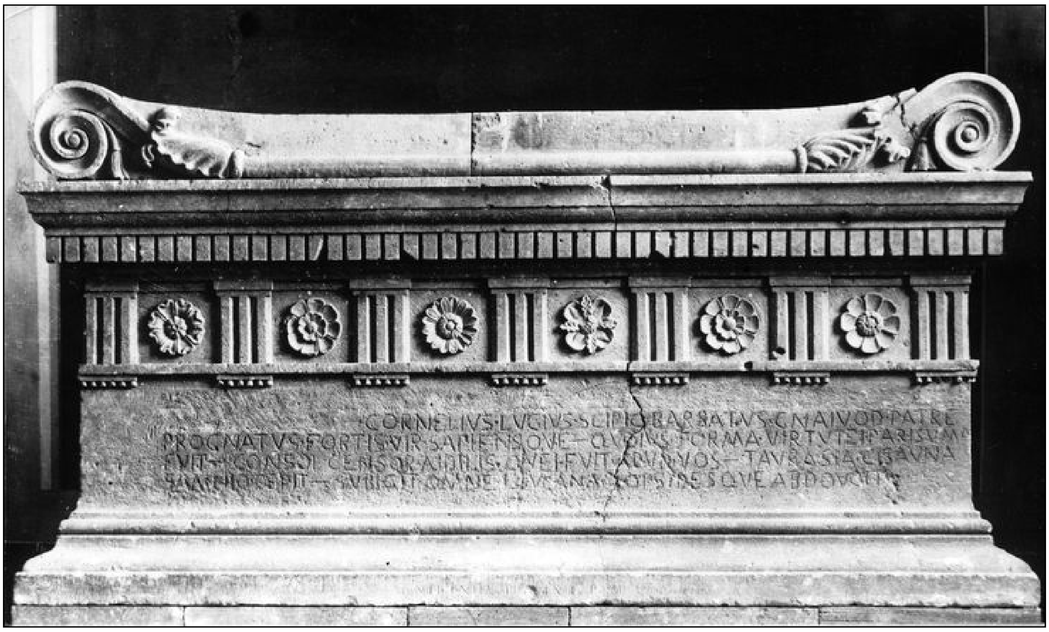


Fig. 3 The sarcophagus of Scipio Barbatus, currently displayed in the Vatican Museums.

In another epigraph composed some thirty years later, the word *pr**ae**fectura* (‘prefecture’) is written *pr**ai**fectura* (*CIL* 10, 06231).[25] Matching the phonetic spellings of these inscriptions with the general linguistic tendency of diphthongs to take as second element a semi-consonantal sound ‘i’ or ‘u’, a past generation of scholars suggested reading classical texts by pronouncing the diphthongs ‘*ae*’ and ‘*oe*’ as /aj/ and /oj/, like the English sounds of ‘high’ and ‘boy.’[26] However, scholars have more recently pointed out that these pronunciations might simply reflect an archaic transition between the diphthong ai and ae, since the canonical form *aedem* is already attested in the famous text of the so-called *senatus consultum de bacchanalibus*, written in 186 BCE.[27] Thus, while maintaining that the second element of a diphthong is always an asyllabic vowel that cannot be stressed, modern scholarship has suggested that, in Classical Latin, “the pronunciation of the diphthongs *ae* and *oe* is [ae] and [oe] respectively.”[28] The introductory section of the app presents the results of these studies in the form of simple Latin rules, often comparing the sounds of Latin with that of modern languages.

LATIN ALPHABET	CLASSICAL PRONOUNCIATION	ECCLESIASTICAL PRONOUNCIATION
ā	aː	a
ă	a	a
(ae)	ae̯	ɛ
b	b	b
c	k	k
c + e, i, y, ae, oe	k	tʃ
ch	kʰ	k
d	d	d
ē	eː	e

[25] *CIL* is an acronym standing for *Corpus Inscriptionum Latinarum*: this work contains a comprehensive collection of ancient Latin inscriptions.

[26] See Allen (1966), 131–32.

[27] See Cupaiuolo (1991), 77–87.

[28] Quotation from Oniga (2014), 22. See also Cupaiuolo (1991), 86–87 and Traina/Bernardi-Perini (1998), 66.

ĕ	ɛ	e
f	f	f
g	g	g
g + e, i, y, ae, oe	g	dʒ
gn	ŋn	ɲ
h	h	-
ĭ	ɪ	i
ī	iː	i
i (semiconsonant)	j	j
k	k	k
l	l	l
m	m	m
n	n	n
ŏ	ɔ	o
ō	oː	o
(oe)	oe̯	e
p	p	p
ph	pʰ	f
q	kʷ	kw
r	r	r
s	s	s
t	t	t
th	tʰ	t
ŭ	ʊ	u
ū	uː	u
u (semiconsonant)	w	v
x	ks	ks
z	z	dz

Tab. 1 Latin phonetic transcription.

The second part of the app's introductory section contains a list of the most common Latin lemmas which are crucial for students in their vocabulary-learning. Ideally, it would have been nice to have a dedicated section for this frequency list. However, toolbars of mobile apps tend to be very limited in terms of space. Therefore, we decided to place the list of Latin lemmas in the introductory section, after the explanation of Latin phonetics. While the online program (in GitHub) can virtually scan every Latin lemma, the app offers a selection of the most common 6500 Latin words as attested across a wide corpus of Classical and Christian texts dating from the fourth century BCE to the sixth century CE.[29] Following a growing scholarly consensus that frequency information plays a key role not only in computational linguistics but also in literary and intertextual research, and in language teaching, the last two

[29] The online Python transcriber can be accessed at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/CLASSICAL%26ECCLESIASTICAL%20LATIN%20IPA%20TRANSCRIBER (last access 02.09.2020). The App frequency list is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/Mobile%20APP/INFO_LIST.txt (last access 02.09.2020).

decades have seen the publication of many frequency dictionaries for modern languages.[30] Yet, while Latin authors themselves often referred to the frequency or usus of Latin words in their commentaries, no comprehensive Latin frequency dictionary exists.[31] The few modern attempts to provide a rigorous lemmatization and count of Latin words have always adopted limited textual corpora based on 'highly representative' authors from the so-called 'golden literature.'[32] These dictionaries consequently struggle to meet the needs of contemporary students and researchers, who are increasingly shifting their attention to the 'less famous' literature of the early Republican, Christian, and Late Antique periods. By contrast, the frequency list provided by the app is based on a wide corpus of 307 Classical and Christian authors, which has been analyzed using original algorithms and the capabilities of the new lemmatizer *Lemlat* to provide a realistic picture of the most common terms used in Latin texts of different periods.[33] Since average cultured speakers of a language know around twenty thousand lemmas, and use only a few thousand of them in their daily life, our 6500-word list provides students not only with basic lemmas, but also with the most important technical and specific words most commonly attested in Latin literature.[34] At the same time, the reasonably small size of the corpus makes it possible for the information provided to be manually checked, and for the app to work even offline.

The section 'search' contains the most important contribution offered by the app: the phonetic transcription of Latin lemmas in both Classical and Ecclesiastical Latin.[35] Using this function, users can type a Latin lemma without diacritics, and access information on the quantities of its syllables, and its accentuation, pronunciation(s), and basic grammatical information, including the presence of homographs that have different meaning and prosody. For example, when one searches the word *praedico*, the program shows that two lemmas have the same spelling, one of them with the penultimate syllable short and being a verb of the first conjugation (*prāedĭco*, *prāedĭcas*, *praedĭcāre*, *prāedicavi*, *prāedicatum*; to announce), while the other has the penultimate syllable long and belongs to the third conjugation (*prāedīco*, *prāedīcis*, *prāedīcĕre*, *prāedīxi*, *prāedīctum*; to foretell). Although the verb *prāedīco* is less common than the verb *prāedĭco*, in these cases, the database displays both entries to help users note potentially ambiguous forms, showing eventual differences in their pronunciation.

To make the program more accessible to beginner students, the app provides not only the IPA transcription, but also the syllabication and accentuation of each lemma using the Latin alphabet. This information, displayed between squared brackets, can be used for both Classical and Ecclesiastical Latin.[36] However, when reading late-antique and medieval texts in Ecclesiastical pronunciation, users should be aware that, after the quantity of vowels was no longer perceived by Latin speakers, the accentuation of some words changed. For instance, while Classical Latin could not preserve the original accentuation of Greek words such as *φιλοσοφία* (philosophy), which was pronounced *philosóphĭa* according to the Latin

[30] On the importance of frequency lists in the pedagogy of Latin, see Muccigrosso (2004). On the use of frequency data for language teaching in general, see Sinclair (1991), 30 and Davies (2005), vii.

[31] See Folco Martinazzoli (1953) on the use of the concept of *hapax legomenon* by ancient commentators and Denooz (2010), 1–2.

[32] Latin frequency dictionaries have been published by Diederich (1939); Delatte/Evrard/Govaerts/Denooz (1981); and Denooz (2010). The largest corpora used so far is that of Denooz (2010), which includes nineteen authors but does not include important texts such as Ovid's *Metamorphoses*.

[33] The original source code is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src (last access 02.09.2020). The lemmatizer *Lemlat* can be accessed at http://www.lemlat3.eu/ (last access 12.01.2021).

[34] On modern languages, see, for instance, Coxhead/Nation/Sim (2015), 121–35. Modern Latin frequency lists tend to provide students with only a few thousand terms. For instance, Williams (2012) offers a 1425 word-list.

[35] The database is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/Mobile%20APP/Data_App_accentuation_Ipa.txt (last access 02.09.2020).

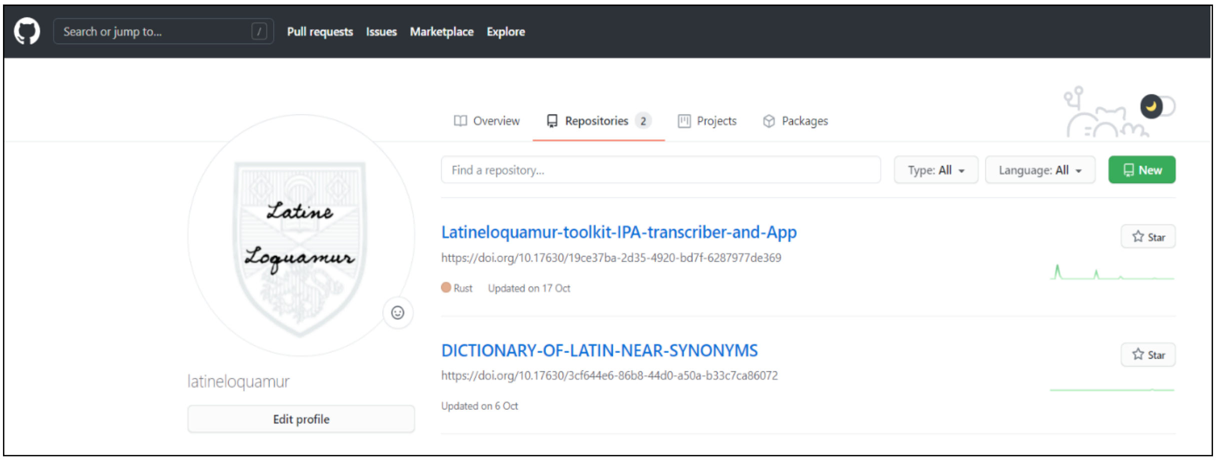
[36] See Pavanetto (2009).

prosody, the Greek words introduced into Latin vocabulary after the disappearance of vocalic quantities around the third century CE maintained their original Greek accentuation (e.g., *ἔρημος*, *éremus*, 'hermitage'). Similarly, the words in which the penultimate syllable was short and was followed by a *muta cum liquida*, which were stressed on the third last syllable in Classical Latin (e.g., *íntĕgrum*; intact), tended to be accentuated on the penultimate syllable in late-antique and medieval Latin (e.g., *intégrum*).[37]

Section Three, which users can access through the button 'Info', explains the genesis of the app and acknowledges the work of the Classicists (T. Spinelli, C. Pavanetto) and Computer Scientists (Giacomo Fenzi, Kamil Kolosowski, Jan Rybojad) who developed it. Moreover, it contains links to the online repositories in which the codes and programs underpinning the app are stored. Overall, in its unique and unprecedented features, the *Handbook of Latin Phonetics* app contributes importantly to language teaching and to stylistic and prosodic studies by allowing even beginner students and non-expert users to learn the most common Latin words and their correct pronunciations, as recommended by the most recent studies on Latin linguistics.

The online toolkit: outline and features

Available in opensource, the *Latin Phonetics* online toolkit contains the source codes through which the data displayed in the app is generated. While the app offers only premade information that can be easily accessed (even offline) by every user who is able to use a smartphone, the online program allows users with good informatic skills to generate customized results. As I have anticipated in the introduction, the online program is accessible through the GitHub page of the *Latine Loquamur Project*.[38] The project's home page currently features two repositories containing, respectively, the *Online Dictionary of Latin Near Synonyms* (which I plan to discuss in another article), and the program on Latin phonetics, which can be accessed by clicking on the folder 'Latineloquamur-toolkit-IPA-transcriber-and-App'.



Latineloquamur-toolkit-IPA-transcriber-and-App
https://doi.org/10.17630/19ce37ba-2d35-4920-bd7f-6287977de369
Rust Updated on 17 Oct

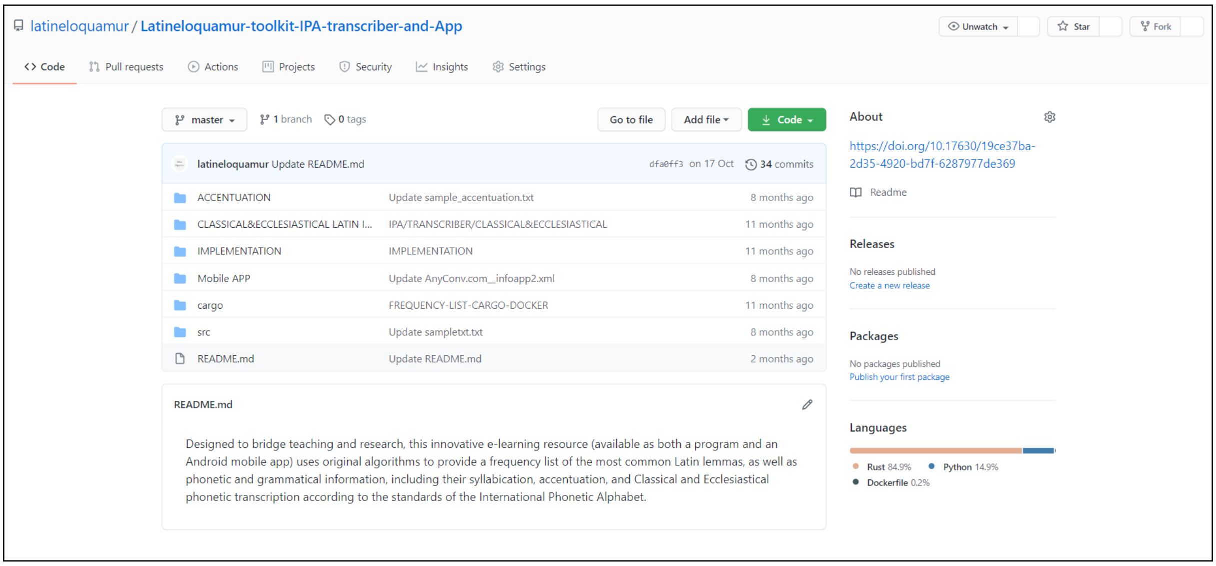
DICTIONARY-OF-LATIN-NEAR-SYNONYMS
https://doi.org/10.17630/3cf644e6-86b8-44d0-a50a-b33c7ca86072
Updated on 6 Oct

Fig. 4 The homepage of the Latine Loquamur repository in GitHub.

The repository that hosts the program on Latin phonetics is organized in different folders corresponding to the different functions performed by the program. This means that, while one can see the accentuation, syllabication, phonetic transcription and potential homographic forms of selected lemmas simultaneously in the app, the online program generates these results separately through different packages.

[37] On the evolution of Latin through Late Antiquity, see Norberg (1999), 33–35.

[38] https://github.com/latineloquamur?tab=repositories (last access 02.09.2020).



latineloquamur Update README.md dfa0ff3 on 17 Oct 34 commits

ACCENTUATION	Update sample_accentuation.txt	8 months ago
CLASSICAL&ECCLESIASTICAL LATIN I...	IPA/TRANSCRIBER/CLASSICAL&ECCLESIASTICAL	11 months ago
IMPLEMENTATION	IMPLEMENTATION	11 months ago
Mobile APP	Update AnyConv.com__infoapp2.xml	8 months ago
cargo	FREQUENCY-LIST-CARGO-DOCKER	11 months ago
src	Update sampletxt.txt	8 months ago
README.md	Update README.md	2 months ago

README.md

Designed to bridge teaching and research, this innovative e-learning resource (available as both a program and an Android mobile app) uses original algorithms to provide a frequency list of the most common Latin lemmas, as well as phonetic and grammatical information, including their syllabication, accentuation, and Classical and Ecclesiastical phonetic transcription according to the standards of the International Phonetic Alphabet.

https://doi.org/10.17630/19ce37ba-2d35-4920-bd7f-6287977de369

No releases published
Create a new release

No packages published
Publish your first package

Rust 84.9% Python 14.9%
Dockerfile 0.2%

Fig. 5 The folders in which the Latin Phonetics online program is organized.

The folder ‘Accentuation’ contains the codes to accentuate automatically macronized Latin words organized in a CSV-file.[39] The output of this program is a new CSV-file displaying both the original word and its accentuated form.[40] The folder ‘Classical and Ecclesiastical Latin IPA Transcriber’ hosts the codes that perform the transcription of given Latin lemmas into phonetic characters according to the standards of the International Phonetic Alphabet.[41] Here the file ‘Readme.md’ provides users with a detailed guide on how to run this complex package. In extreme synthesis, by using in sequence the commands *cargo build* and *cargorun--{path to the file}* users can operate the phonetic transcription of Latin words (organized one per line in a txt file) and generate two files containing, respectively, the Classical and Ecclesiastical pronunciation of those words. A sample of the results that can be achieved using this package is provided by the section ‘sample IPA’.[42] The folders ‘Implementation’ and ‘Mobile App’ can be disregarded by users as they contain, respectively, work-in-progress material that will be used in the implementation of the toolkit (as described in the final section of this article) and the codes that have been used to build the app. Using the folders ‘Cargo’ and ‘Src’ expert users can generate frequency lists of Latin lemmas attested in a customizable set of Latin texts.[43] Specifically, the folder ‘cargo’ governs the functioning of the packages hosted in ‘Src’ and contains a ‘Dockerfile’ with instructions. To use this program, users can upload the texts that they wish to process in the sub-folder ‘data_dir’ (within ‘Src’).[44] Files in this repository must be organized in folders, one for each author, and named with the authors’ names. Inside each author-folder, there must be a list of folders corresponding to the author’s works, which must be stored in ‘txt’ format. An example of how to organize personalized textual corpora effi-

[39] Both functions are available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/ACCENTUATION/Latin_accentuation_code.py (last access 02.09.2020).

[40] A sample is available at (https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/ACCENTUATION/sample_accentuation.txt (last access 02.09.2020).

[41] https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/CLASSICAL%26ECCLESIASTICAL%20LATIN%20IPA%20TRANSCRIBER (last access 02.09.2020).

[42] https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/CLASSICAL%26ECCLESIASTICAL%20LATIN%20IPA%20TRANSCRIBER/sample%20IPA (last access 02.09.2020). Note that slight differences may be caused by the manual checking operated on the data used in the app.

[43] https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/cargo; https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src (last access 02.09.2020).

[44] The repository is accessible at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src/Data_dir (last access 02.09.2020).

ciently is offered by the file ‘sampletxt.txt’.[45] While the Android app simply provides a list of the 6500 most common Latin lemmas, expert users can perform much more complex searches with the online program to generate data on the use of a term or only on some of its forms by specific authors within a selected time frame. The tremendous potential of this tool can be appreciated by looking at the sample that, stored in the GitHub repository, shows a part of the data generated by running the program on our large textual corpus.[46]

Technologies

The original technologies used to develop the program underpinning the app are highly advanced and closely tailored to its aims and function. This complex program is organized in different packages that govern, respectively, the frequential statistics of the words most commonly attested in Latin literature, the syllabication and accentuation of Latin lemmas, their transcription into the characters of the International Phonetic Alphabet, and their visualization through a user-friendly mobile app.[47] The following section will discuss the most important technologies and methodological issues concerning each component of the backend.

Lemmatizing and counting Latin

The first stage of the development of the *Latin Phonetics* toolkit was the creation of a unique frequency list of the 6500 most common Latin lemmas, as attested across a large corpus of both Classical and Christian texts. Making this list involved parsing, lemmatizing, and categorizing the data directly from the sources, which means scanning a pre-assembled and standardized textual corpus in order to identify the different inflected forms of each word, and to calculate which lemmas are the ones most frequently used by Latin authors.

While most previous Latin dictionaries have relied on a manual processing of texts, this toolkit uses original algorithms and the capabilities of the opensource lemmatization service offered by *Lemlat*.[48] Lemmatization is the process through which the variants of a term, and its inflected or graphically different forms (e.g., *amat*, *amant*, *amas*, *amavi*, *amatum*; to love), are attributed to their lemma: the standard form of the word (e.g., *amo*) as it appears in a dictionary. Many programs can perform this task on Latin texts quite successfully (e.g., the Schinke algorithm, the Perseus lemmatizer, *PROIEL*, *Parsley*, *Morpheus*, Whitaker’s *Words*, *LatMor*), but none of these technologies provides entirely correct data.[49] Among them, we have chosen to use a freely adapted version of *Lemlat*, which was developed between 2002 and 2004 by the National Research Centre (CNR) of Pisa in collaboration with the University of

[45] https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/commit/b475af49a9dcbabb3a9cb-70582da84b5df18ecd9 (last access 13.12.2020).

[46] Samples are available on Github (https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/src/sample_frequency_list_data.txt; last access 02.09.2020) and in a dedicated, work-in-progress webpage (https://latin.netlify.com/; last access 02.09.2020).

[47] These packages are freely accessible through the program’s repository: https://doi.org/10.17630/19ce37ba-2d35-4920-bd7f-6287977de369 (last access 02.09.2020); https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App (last access 02.09.2020).

[48] Cf. http://www.ilc.cnr.it/lemlat/lemlat/index.html (last access 12.01.2021).

[49] See, for instance, *LatMor* (http://cistern.cis.lmu.de; last access 02.09.2020), *Words* (http://archives.nd.edu/words.html; last access 02.09.2020), *Parsley* (https://github.com/goldibex; last access 02.09.2020), *PROIEL* (https://github.com/proiel/proiel-treebank; last access 02.09.2020), and *Morpheus* (https://github.com/tmallon/morpheus; last access 02.09.2020). On these technologies see also Springmann/Schmid/Dietmar (2016).

Turin, because it has proved to be the most consistent and reliable technology of this kind.[50] Based on a database of 40,014 lexical entries and 43,432 lemmas including many late antique and medieval terms, *Lemlat* adopts the standards of the *Oxford Latin Dictionary* (Glare [1982]). Being able to recognize over 97% of Latin terms including many anthroponyms and toponyms, it successfully lemmatizes 319,725 lexemes into 30,413 lexical entries (around 3,500 more than the modern Liège dictionary by Denooz). Moreover, its automatic analysis is very accurate and takes into account many spelling variations and even rare or archaic forms of a lemma which the former frequency dictionaries neglect.[51] However, this technology, which is still undergoing further development, cannot disambiguate homographic forms, which are therefore counted under all the lemmas which they can belong to.

To create the frequency list used in the app, we have fed into the program (written in RUST) a large textual corpus yielding some 9,484,029 words, and covering the works of 307 authors, which has been created using different opensource textual databases available online such as Perseus, the PHI database, and the *Bibliotheca Augustana*.[52] This textual corpus, which has not been made publicly available in accordance with its distribution licence, was stored in the repository *data_dir*.[53] As I have anticipated, this folder has been left empty in the program's repository, so that users can input a personalized corpus on which they can run our program by using, for example, the big textual databank provided by *Perseus*, or the *Packard Humanities Institute* both online and on CD. The most important element of this package is the 'lemmatizer.' This file is a 'CSV' directly exported (with adaptations) from Lemlat to specify the lemmatization bases that are used to operate on the literature.[54] This section also contains an original 'runner' program that is in charge of the full *GraphQL* endpoint, being used to query the text through the generic command: '*cargo run--bin {program_name}---aAUTHORS_FILE-dDATA_DIR-lLEMM_FILE.*' There, the options *authors_file*, *data_dir*, *lemm_file* are used to specify the data files on which to operate. Specifically, *authors_file* is used for an advanced function which is still being perfected; this folder contains a database with the chronology of Latin authors which can be used to perform

[50] The program is available at http://www.lemlat3.eu/ (last access 02.09.2020). On its features and assessment see Passarotti/Baudassi/Litta/Ruffolo (2017), 24–31 and Springmann/Schmid/Dietmar (2016). The adapted version of the lemmatizer is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src/latin_lemmatizer (last access 02.09.2020).

[51] See Passarotti/Baudassi/Litta/Ruffolo (2017), 26. A way to check the ways in which forms are lemmatized in our dictionary through *LEMLAT* is through http://www.ilc.cnr.it/lemlat/lemlat/index.html (last access 12.01.2021). The lemmatizer Lemlat has successfully lemmatized more than 97% of the word-forms attested in our corpus, leaving unrecognized only the 2.88% of the forms. Among them many are names, Greek forms used in Latin texts, indication of books given in Roman letters (e.g., LXV), or Latin endings (e.g., *-ar*; *-or*) that are mentioned by ancient grammarians in their discussions of Latin morphology but do not correspond to any lemma.

[52] The RUST source code is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src (last access 02.09.2020). On *Perseus* see http://www.perseus.tufts.edu/hopper/ (last access 02.09.2020); the Bibliotheca Augustana can be accessed at http://www.hs-augsburg.de/~harsch/a_impressum.html (last access 02.09.2020). The list of authors included in our textual corpus is stored in this repository: https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/src/authors_chrono/AUTHORS-LIST (last access 02.09.2020). Our corpus uses the *PHI* standards to name Latin authors in order to facilitate the use of the toolkit by other users who will likely use the *PHI* textual database. The *Packard Humanities Institute* (*PHI*) corpus is one of the widest opensource Latin corpora currently available online https://latin.packhum.org/ (last access 02.09.2020) and, although it does not match our corpus perfectly, it can be effectively used to look up the large majority of the Latin passages in which our lemmas or their inflected forms appear. However, while our program lemmatizes every inflected form, the *PHI* searching tool performs only simple pattern-matching queries. Thus, if one searches 'ultor' the program shows also results like '***multorum***', unless the search is made for a specific form like #*ultor*#. In this case the program displays only the occurrences of this specific graphic form and not of the lemma *ultor* and of its inflected forms.

[53] The repository is accessible at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src/Data_dir (last access 02.09.2020).

[54] To run the lemmatizer use https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src/latin_lemmatizer/src/parsers (last access 02.09.2020).

diachronic linguistic searches on specific centuries.[55] The data so created can be exported using the *CSV* function. Conceptually, this system works as a 'forest' of sorts, with one tree for each lemma, one leaf for each form, and each form with a collection of occurrences. The data obtained from this system is semi-structured (data is fully tagged, but its structure is not rigidly defined, allowing for great flexibility in terms of exporting it), and the entire system operates without interacting with the disk. Leveraging this core system, two applications can query the relational databases (texts, lemmatizer, and chronological authors map), and generate the results requested by users in different formats (e.g., Json, txt, Excel).[56]

Prosodic processing

After building the list of the most common Latin lemmas, other algorithms were used to divide words into their syllables, and to mark them as long or short respectively, which is indispensable for a correct phonetic transcription. In Latin, the quantity of syllables does not always coincide with the quantity of the vowels that they contain. However, syllables are always long, except where a short vowel is in an open syllable (a syllable that does not end with a consonant). For instance, the *u* in the second syllable of the word *sepultus* (buried) is short by nature (**se-pŭl-tus*). However, because it is closed (it ends in consonant), this syllable is long, and takes the accent (*sepúltus*). For this reason, the program displays the quantities of syllables. Several opensource tools can divide Latin words into syllables, and mark the long ones as such.[57] Among them, we used the syllabifier shared by CLTK and Collatinus because it has been recently implemented to correctly process 'exceptional' forms that do not follow the standard rule of syllabication, using a list made by Rev. Frère Romain Marie de l' Abbaye Saint-Joseph de Flavigny-sur-Ozerain in 2016.[58] Thus, this algorithm can correctly syllabify compound words in which consonants are counted in the same syllable (e.g., *de-scri-bo*; to describe). This tool also offers the most efficient macronizer currently available, which is based on eight Latin dictionaries.[59]

After marking the quantities of each syllable, original Python scripts were used to accentuate Latin lemmas.[60] This program, which I co-developed in collaboration with Jan Rybojad, takes as an input a CSV-file containing Latin lemmas, and parses words so as to break them into an array of Unicode characters. For each lemma, this array is further converted into two new arrays, one for sounds, and one for vowels (including diphthongs). The actual accentuation is performed through the functions *findStress* and *isLongVowel* that replace long vowels and diphthongs with the appropriate stressed vowels and diphthongs, according to the rules of Latin accentuation.[61] In particular, if the second-last vowel of a lemma is marked as long or is a diphthong, the program accentuates it; if the second-last vowel is marked as short

[55] The runner program is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/src/runner (last access 02.09.2020).

[56] Samples of potential results are available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/src/sample_frequency_list_data.txt; last access 02.09.2020) and https://latin.netlify.com/ (last access 0.09.2020).

[57] For instance, this service is provided by: http://marello.org/tools/syllabifier/ (last access 02.09.2020); https://github.com/cltk/cltk/blob/master/cltk/stem/latin/syllabifier.py (last access 02.09.2020); https://github.com/biblissima/collatinus/blob/master/doc-usr/scander.md (last access 02.09.2020).

[58] Cf. https://github.com/biblissima/collatinus/blob/master/bin/data/hyphen.la (last access 02.09.2020).

[59] The dictionaries are De Valbuena (1819); Noël (1824); Quicherat (1836); De Miguel (1867); Franklin (1875); Lewis/Short (1879); Du Cange (1883); Georges (1888); Calonghi (1898); Gaffiot (1934); Gaffiot (2016).

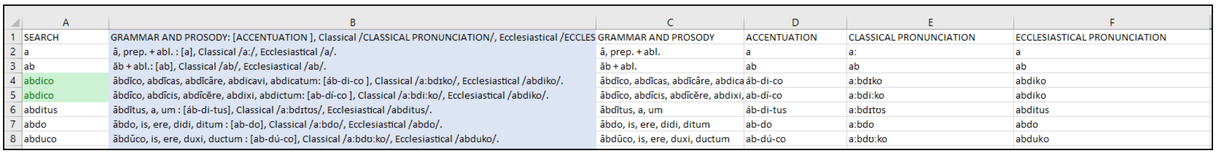
[60] See https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/ACCENTUATION (last access 02.09.2020).

[61] Both functions are available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/ACCENTUATION/Latin_accentuation_code.py (last access 02.09.2020).

and not a diphthong, then it replaces the third-last vowel or diphthong with the corresponding stressed vowel. The output is a new CSV-file with lines in the format <word>, <accentuated word>.[62]

IPA transcriber for Classical and Ecclesiastical Latin

A unique feature of the *Latin Phonetics* toolkit is the phonetic transcription of both Classical and Ecclesiastical Latin using the *International Phonetic Alphabet*, which is performed by original algorithms. Because the program takes as input Latin words that have the quantities of their syllables fully marked, this operation is conceptually simple. Given an input word, the algorithm applies iteratively a number of replacement rules that, co-developed in collaboration with Giacomo Fenzi, convert a combination of Latin characters into the corresponding IPA symbols (e.g., *x* → /*ks*/). However, several factors make this process more complex. Firstly, the pronunciation of Classical and Ecclesiastical Latin has to be treated separately because it follows different rules. For instance, while the long e (*ē*) is pronounced as long /e:/ in Classical Latin, it is pronounced as a normal closed /e/ in Ecclesiastical Latin, where the quantity of vowels is no longer perceived as a phonetically significant element. Similarly, the nexus *gn*, which sounds /ŋn/ in Classical Latin, is softened in Ecclesiastical Latin (/ɲ/). In this important respect, the phonetic transcription of Classical Latin operated by the *Latin Phonetics* toolkit differs from that of CLTK in so far as it is based on the new phonetic transcriptions recommended by recent studies of Latin linguistics. Secondly, combinations of letters are sometimes pronounced as just one sound. While a *nexus* can have different lengths, the same letters that appear in a two-character group can be pronounced differently when they occur in a three-character nexus. For instance, in the term *ămīcĭtĭa* (friendship), the nexus '*tia*' is pronounced /tsja/ in Ecclesiastical Latin. However, in the plural form *ămīcĭtĭāe* the same group '*tia*' appears in the longer group '*tiae*' which is formed by the nexus '*ti*+vowel' and the diphthong '*ae*'. In this case, the replacement /tsja/+/e/ would be wrong, because the group *ae* is monophthonged in /ɛ/ or /e/ in Ecclesiastical Latin, and the word is consequently pronounced /a.miˈtʃi.tsje/. To fix these problems, the program, which is written in RUST, processes Classical and Ecclesiastical Latin separately.[63] In particular, using as an input a path to a file containing a list of words (one per line), the functions 'cargo build' (the executable being 'target/debug/ipa_latin(.exe)') and 'cargorun--{path to the file}' operate parallel replacement for Classical and Ecclesiastical Latin, using strings such as 'if Classical {subs.push((„aei", „ae̯i"));} else {subs.push((„aei", „ɛi"));}', where *else* is the Ecclesiastical pronunciation. As a result, the program generates two different files: 'Classical.txt' and 'Eccl.txt'. In order to efficiently treat *nexus*, such replacements are based on conversion rules that, specifying each possible combination, are applied in descending length-order, so as to match longer structures first.[64] In this way, for instance, the group *oe* is successfully transcribed as /e/ in Ecclesiastical Latin, rather than as /o/+/e/. The results of this process can be seen in the files stored in the GitHub folder 'sample IPA'.[65]



	A	B	C	D	E	F
1	SEARCH	GRAMMAR AND PROSODY: [ACCENTUATION], Classical /CLASSICAL PRONUNCIATION/, Ecclesiastical /ECCLES	GRAMMAR AND PROSODY	ACCENTUATION	CLASSICAL PRONUNCIATION	ECCLESIASTICAL PRONUNCIATION
2	a	ā, prep. + abl. : [a], Classical /a:/, Ecclesiastical /a/.	ā, prep. + abl.	a	a:	a
3	ab	ăb + abl.: [ab], Classical /ab/, Ecclesiastical /ab/.	ăb + abl.	ab	ab	ab
4	abdico	ābdĭco, abdĭcas, abdĭcāre, abdicavi, abdicatum: [áb-di-co], Classical /a:bdɪko/, Ecclesiastical /abdiko/.	ābdĭco, abdĭcas, abdĭcāre, abdica	áb-di-co	a:bdɪko	abdiko
5	abdico	ābdīco, abdīcis, abdīcĕre, abdixi, abdictum: [ab-dí-co], Classical /a:bdi:ko/, Ecclesiastical /abdiko/.	ābdīco, abdīcis, abdīcĕre, abdixi,	ab-dí-co	a:bdi:ko	abdiko
6	abditus	ābdĭtus, a, um : [áb-di-tus], Classical /a:bdɪtʊs/, Ecclesiastical /abditus/.	ābdĭtus, a, um	áb-di-tus	a:bdɪtʊs	abditus
7	abdo	ābdo, is, ere, didi, ditum : [ab-do], Classical /a:bdo/, Ecclesiastical /abdo/.	ābdo, is, ere, didi, ditum	ab-do	a:bdo	abdo
8	abduco	ābdūco, is, ere, duxi, ductum : [ab-dú-co], Classical /a:bdʊ:ko/, Ecclesiastical /abduko/.	ābdūco, is, ere, duxi, ductum	ab-dú-co	a:bdʊ:ko	abduko

Fig. 6 Sample of the database deployed by the app.

[62] A sample is available at (https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/ACCENTUATION/sample_accentuation.txt (last access 02.09.2020).

[63] The RUST transcriber is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/CLASSICAL%26ECCLESIASTICAL%20LATIN%20IPA%20TRANSCRIBER (last access 02.09.2020).

[64] https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/CLASSICAL%26ECCLESIASTICAL%20LATIN%20IPA%20TRANSCRIBER/src/ipa.rs (last access 02.09.2020).

[65] https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/tree/master/CLASSICAL%26ECCLESIASTICAL%20LATIN%20IPA%20TRANSCRIBER/sample%20IPA (last access 02.09.2020). Note that slight differences may be caused by the manual checking operated on the data used in the app.

App design

The project has also pioneered the creation of an intuitive mobile app that, titled *Handbook of Latin Phonetics*, makes the phonetic program easily accessible for students and non-expert users. Because the amount of data is relatively small, rather than connecting the app to the program through a rest API, the application has been designed to work offline, using as input a SQLite database containing the information created by the program. In this file, a list of Latin lemmas without diacritics is matched with the lemmas complete with grammatical and prosodic information and their phonetic transcriptions. Therefore, when one searches a word without diacritics, all the possible corresponding Latin lemmas, which may include different syllabic quantities as in the aforementioned case of *praedico*, are returned. The app was built using a Software Development Kit called *Flutter*, which allows one to build high-performance apps for iOS & Android from a single codebase, using Dart programming language.[66] In the future, it will become the native framework of Google's Fuchsia *OS*, so that a project developed in Flutter will work on three platforms: iOS, Android, and Fuchsia. The architecture of the app is simple, and uses a Business Logic Components pattern, meaning that everything in the app is represented as a stream of events, in which widgets submit events and other widgets respond.

Future developments

Two new implementations of the *Latin Phonetics* toolkit are being developed to further support both academic research and language teaching.[67] The first is a diachronic function which, based on an original diachronic mapping of Latin authors, will allow users to see by which Latin authors and in which century each lemma was used.[68] This feature will support not only stylistic choices in exercises of Latin composition, but also commentary writers (by providing a concise 'story' of each word and of its occurrences) and philological conjectures (by showing which words were more likely to be used by an author). The second new feature that is being designed will leverage my app of Latin synonyms to describe the meaning of each lemma (for which phonetic information is provided) directly in Latin through the list of its most important Latin synonyms.[69] This function will also support new digital technologies that are being developed to detect similarities of meanings and ideas between Latin texts, independently of precise lexical repetitions. Finally, a version of the app for Apple devices will be released soon.

Overall, in its innovative cross-fertilization of recent developments in the fields of Latin linguistics, pedagogy, and digital humanities, the *Latin Phonetics* toolkit bridges teaching and research, using original algorithms to provide scholars and students with the first IPA phonetic transcription of the Classical and Ecclesiastical pronunciations of the most common Latin lemmas, as attested across the entire corpus of Latin literature. Besides facilitating the teaching of Latin in Latin and contributing to the creation of a shared methodology for the study of Latin phonology, this tool also supports a more interactive inde-

[66] An overview of this innovative technology is provided by Kuzmin/Ignatiev/Grafov (2020). Cf. https://flutter.dev/ (last access 02.09.2020).

[67] Originally, we had planned to assess the impact of our toolkit (published at the beginning of 2020) by using students' and professors' feedback. However, due to the Covid-19 pandemic this has been impossible so far. We now aim to collect and examine feedback after the development of the two new implementations.

[68] A draft is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/src/authors_chrono/AUTHORS-LIST (last access 02.09.2020).

[69] A sample is available at https://github.com/latineloquamur/Latineloquamur-toolkit-IPA-transcriber-and-App/blob/master/IMPLEMENTATION/IMPLEMENTATION_SYNONYMS.txt (last access 02.09.2020).

pendent learning of Latin, displaying the benefits of truly interdisciplinary approaches to the study of classical languages.[70]

[70] I would like to express my sincere gratitude to the University of St Andrews and the *Pontificium Institutum Altioris Latinitatis* (Salesian University of Rome) for the generous support that they provided to my project; to William Short, Cleto Pavanetto and Miran Sajovic, who helped me perfect several aspects of my research; to Giacomo Fenzi, Kamil Kolosowski and Jan Rybojad who helped me develop the digital program and the app; to Gregory Tirenin and Maxwell Stocker who proofread this article.

References

Allen (1966): W. S. Allen, Vox Latina: a guide to the pronunciation of Classical Latin, Cambridge 1966.

Avitus (2018): A. G. Avitus, Spoken Latin: Learning, Teaching, Lecturing and Research, Journal of Classics Teaching 19.37 (2018), 46–52.

Barnett (2006): F. J. Barnett, The second appendix to Probus, The Classical Quarterly 56.1 (2006), 257–278.

Calonghi (1898): F. Calonghi, Dizionario Latino-Italiano, Milan 1898.

Chiesa (2012): P. Chiesa, L‘impiego del ‚cursus‘ in sede di critica testuale: una prospettiva diagnostica, in: F. Bognini, Meminisse iuuat. Studi in memoria di Violetta De Angelis, Florence 2012, 279–304.

CIL = *Corpus Inscriptionum Latinarum*, ed. Th. Mommsen et alii, Berlin 1862ff.

Collins (2012): A. Collins, The English pronunciation of Latin: its rise and fall, Cambridge Classical Journal, 58 (2012), 23–57.

Collins (1988): J. F. Collins, A primer of Ecclesiastical Latin, Washington 1988.

Coxhead/Nation/Sim (2015): A. Coxhead/ P. Nation/ D. Sim, Measuring the Vocabulary Size of Native Speakers of English in New Zealand Secondary Schools, NZ. J. Educ. Stud. 50 (2015), 121–135.

Cupaiuolo (1991): F. Cupaiuolo, Problemi di lingua latina. Appunti di grammatica storica, Naples 1991.

Davies (2005): M. Davies, A frequency Dictionary of Spanish: Core vocabulary for learners, Abingdon 2005.

De Miguel (1867): R. De Miguel, Nuevo Diccionario Latino-Español Etimológico, Leipzig 1867.

De Valbuena (1819): M. De Valbuena, Diccionario Universal Latino-Español, Hermanos 1819.

Delatte/Evrard/Govaerts/Denooz (1981): L. Delatte, É. Evrard, S. Govaerts, J. Denooz, Dictionnaire fréquentiel et index inverse de la langue latine, Liege 1981.

Denooz (2010): J. Denooz, Nouveau lexique fréquentiel de latin. Alpha-Omega, Liege 2010.

Diederich (1939): P. B. Diederich, The Frequency of Latin Words and their Endings, Chicago 1939.

Du Cange (1883): C. Du Cange, Glossarium Mediae et Infimae Latinitatis. Ed. L. Favre, Niort 1883–1887.

Franklin (1875): A. Franklin, Dictionnaire des noms, surnoms et pseudonymes latins de l’histoire littéraire du Moyen Age, Paris, 1875.

Gaffiot (1934): D. L. F. Gaffiot, Dictionnaire Latin-Français, Paris 1934.

Gaffiot (2016): D. L. F. Gaffiot, Dictionnaire Latin-Français, London 2016.

Georges (1888): K. E. Georges, Kleines deutsch-lateinisches Handwörterbuch, Hannover / Leipzig 1888.

Harrington/Pucci (1997): K. P. Harrington, J. Pucci, (eds.), Medieval Latin, Chicago 1997.

Kuzmin/Ignatiev/Grafov (2020): N. Kuzmin, K. Ignatiev, D. Grafov, Experience of Developing a Mobile Application Using Flutter. Lecture Notes in Electrical Engineering 621 (2020), 571–75.

Lewis/Short (1879): C. T. Lewis, C. Short, Latin Dictionary, founded on Andrews' Edition of Freund's Latin Dictionary: Revised, Enlarged, and in Great Part Rewritten, Oxford 1879.

Martinazzoli (1953): F. Martinazzoli, Hapax legomenon: Parte prima, Roma 1953.

Muccigrosso (2004): J. D. Muccigrosso, Frequent vocabulary in Latin instruction, The Classical World 97.4 (2004), 409–433.

Noel (1822): F. Noel, Dictionarium Latino-Gallicum: dictionnaire latin-français, compose sur le plan de l'ouvrage intitulé: Magnum totius latinitatis lexicon, de Facciolati, septième edittion, Paris 1822.

Norberg/Oldoni (1999): D. Norberg, M. Oldoni, (eds), Manuale di Latino Medievale, Napoli 1999.

Oniga (2003): R. Oniga, La sopravvivenza di lingue diverse dal latino nell'Italia di età Imperiale: alcune testimonianze letterarie, Lexis: poética, retórica e comunicaciones nella tradizione classica, 21 (2003), 39–62.

Oniga (2014): R. Oniga, Latin: a linguistic introduction, Oxford 2014.

Passarotti/Baudassi/Litta/Ruffolo (2017): M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The 'Lemlat3.0' Package for Morphological Analysis of Latin, Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, 133 (2017), 24–31.

Paulus VI (1976): Paulus P. P. VI, Romani Sermonis, Roma 1976.

Pavanetto (2009): C. Pavanetto, Elementa Linguae et Grammaticae Latinae, sexta editio aucta et emendata, Roma 2009.

Préaux (1957): J. Préaux, Premier congrès international pour le latin vivant: Avignon 3–6 septembre 1956, Latomus, 16.3 (1957), 509–511.

Quicherat (1836): L. Quicherat, Dictionnaire Français-Latin, Paris 1836.

Ramage (1963): E. S. Ramage, Urbanitas: Cicero and Quintilian, a contrast in attitudes, The American Journal of Philology, 84.4 (1963), 390–414.

Serianni (1998): L. Serianni, Lezioni di grammatica storica italiana, Roma 1998.

Short/George (2013): E. Short, A. George, A Primer of botanical Latin with vocabulary, Cambridge 2013.

Sinclair (1991): J. Sinclair, Corpus, concordance, collocation. Oxford (1991).

Spinazzè (2014): L. Spinazzè, Cursus in clausula. An Online Analysis Tool of Latin Prose, Association for Computing Machinery, 10 (2014), 1–6.

Springmann/Schmid/Dietmar (2016): U. Springmann, H. Schmid, N. Dietmar, LatMor: A Latin finite-state morphology encoding vowel quantity, Open Linguistics – Topical Issue on Treebanking and Ancient Languages: Current and Prospective Research, 2.1 (2016), 386–392.

Traina/Bernardi-Perini (1998): A. Traina, G. Bernardi-Perini, Propedeutica al latino universitario, Rome 1998.

Williams (2012): M. A. Williams, Essential Latin Vocabulary: The 1,425 Most Commom Words Occurring in the Actual Writings of Over 200 Latin Authors, Milan 2012.

Winge (2015): J. Winge, Automatic Annotation of Latin Vowel Length. Bachelor’s Thesis in Language Technology, Uppsala University, https://cl.lingfil.uu.se/exarb/arch/winge2015.pdf (last access 02.09.2020).

Figure references

Fig. 1: Latin Phonetics App, by Tommaso Spinelli.
Fig. 2: App’s introductory section, by Tommaso Spinelli.
Fig. 3: Sarcophagus of Scipio Barbatus (Vatican Museums), image by the Center for Epigraphical Studies of the Ohio State University.
(http://db.edcs.eu/epigr/bilder.php?s_language=de&bild=$OH_CIL_06_01284_1.jpg;$OH_CIL_06_01284_2.jpg;$OH_CIL_06_01284_3.jpg;$CIL_01_00006.jpg;PH0010886;PH0010887;pp; last access 12.01.2021).
Fig. 4: The homepage of the *Latine Loquamur* repository in GitHub, by Tommaso Spinelli.
Fig. 5: The folders in which the *Latin Phonetics* online program is organized, by Tommaso Spinelli.
Fig. 6: Database of Latin phonetics, by Tommaso Spinelli.
Tab. 1: Phonetic Transcription Table, by Tommaso Spinelli.

Author contact information[71]

Dr Tommaso Spinelli

University of Manchester
School of Arts, Languages & Cultures

E-Mail: tommaso.spinelli@manchester.ac.uk

[71] The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY 4.0.

##DIGITAL CLASSICS ONLINE##

Building a Repository of Exercises for Learning Latin

Konstantin Schulz

Abstract: This study introduces quality criteria and a reference implementation of an exercise repository for Latin language exercises, with a special focus on vocabulary. The repository is supposed to be easily accessible to people with no prior knowledge of corpus linguistics or natural language processing. Teachers in high schools can generate exercises themselves, which should be fully customizable and adaptive with regard to the learners.
To facilitate the process of creating new exercises from ancient texts, additional linguistic information is needed. For instance, a Keyword-In-Context analysis enables teachers to investigate usage patterns for single words by looking at visualizations of morphological, syntactic and lexical phenomena.
Besides, exercises need to be findable and accessible. To achieve this, a public repository with an underlying database was created, so exercises can be stored and queried according to their relevant metadata, e.g., vocabulary, textual complexity or interaction type. The repository can be used by teachers to retrieve, modify and try out exercises developed by their fellow pedagogues. In this way, didactic efforts can be shared and built upon not just within the same school, but within the whole country, in many cases even internationally.

Introduction

This study seeks to define quality criteria and develop a reference implementation for a repository of language exercises. The exercises were used for research funded by the German Research Foundation in the CALLIDUS[1] project led by Malte Dreyer, Stefan Kipf and Anke Lüdeling. Consequently, the repository is tightly integrated into the project‘s software Machina Callida.[2] This application serves as a combined platform for various tools that help teachers of Latin. Previous publications have highlighted other parts of the platform, such as the generation of corpus-based exercises for learning management systems,[3] or using artificial intelligence to support this curation process.[4]

The present study builds on these papers to introduce novel functionality that has not been described previously: a searchable repository with extensive metadata on born-digital, interactive exercises for learning Latin.[5] The aim is to make existing curated materials accessible to a broader public by allowing people to filter them by, e.g., author, work or interaction type. The selection of relevant metadata needs

[1] Project number 316618374, accessible at https://hu.berlin/callidus (Last access 11.01.2021).

[2] Accessible at https://korpling.org/mc/ (Last access 11.01.2021, public server) and https://scm.cms.hu-berlin.de/callidus/machina-callida (Last access 11.01.2021, source code).

[3] Beyer / Schulz (2020).

[4] Schulz et al. (2020).

[5] The data model was specified using OpenAPI (https://swagger.io/specification/ [Last access 11.01.2021]) and is available at https://scm.cms.hu-berlin.de/callidus/machina-callida/-/blob/master/mc_backend/openapi_models.yaml#L119-188 (Last access 11.01.2021).

to be reflected systematically, which is why the following chapters will make ample use of research literature on corpus linguistics, second language acquisition (esp. Latin pedagogy) and software development. The perspective is therefore interdisciplinary.

Technically, the exercise data is hosted in the cloud services of Humboldt University in Berlin and maintained by the CALLIDUS project. However, access to the data is granted through a public interface, which is represented by a REST API,[6] and by a user-friendly web application.[7] Using the REST API, users can create new exercises and share them with others, thereby contributing to a digital community of Latin teachers. We believe that this combination of reliable technical infrastructure, strict corpus-based approach to language acquisition, and emphasis of collaboration, is a unique characteristic of our learning platform for Latin pedagogy.

In the following, a short analysis of shortcomings in current Latin pedagogy will pave the way for a general estimation of quality criteria for Latin vocabulary exercises. After that, a closer look at evaluation and feedback will reveal important challenges for our system. Finally, we will summarize the major features of our implementation and point to current weaknesses that need to be addressed by future research.

Current State of Learning Latin Vocabulary

In the context of German high schools, various stakeholders are involved in the quality assurance of teaching Latin: students, teachers, researchers, (textbook) publishers and many more. During the last decades, there have been a few efforts to optimize the basic vocabulary for learners at various degrees of proficiency,[8] e.g., by reducing the amount[9] or improving the selection of words to be learned.[10] This discussion about core vocabularies is certainly inspired by the well-known phenomenon of students not being able to understand authentic Latin literature after the first few years of language learning.[11] Such problems arise out of deficient vocabulary training, as can be seen from textbooks focusing on single words instead of meaningful contexts,[12] and repeating important terms too rarely.[13] Therefore, this paper seeks to investigate quality criteria for Latin vocabulary exercises and ways to implement them prototypically in a digital learning environment.

Current vocabularies rely heavily on cross-lingual word equations,[14] which may be suitable for learners at the very beginning,[15] but not at the intermediate and advanced stages because, by this point, students need to develop semantic connections in their mental lexicon.[16] Otherwise, they cannot deal with im-

[6] Accessible at https://korpling.org/mc-service/mc/api/v1.0/ui/ (Last access 11.01.2021).

[7] Accessible at https://korpling.org/mc/exercise-list (Last access 11.01.2021).

[8] Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (2004), 10; Robillard et al. (2014), 2.

[9] Schirok (2010), 17.

[10] Utz (2000), 146.

[11] Schibel (2013), 115.

[12] Waiblinger (1998), 13; Siebel (2011), 127.

[13] Van de Loo (2016), 136.

[14] Van de Loo (2016), 140.

[15] Crossley et al. (2010), 56.

[16] Crossley et al. (2010), 70.

portant aspects of lexical competence, such as collocations, synonymy or derivation,[17] the latter of which has been identified as particularly important for teaching Latin.[18]

Still, many examples of Latin vocabulary software succumb to the temptation of offering decontextualized equations of cross-lingual form-meaning links: This is true for the ‚Vocabulary Drill' in the Wheelock Latin Exercises,[19] for the ‚Vocabulary Handouts' at The Latin Library,[20] and for flashcard applications like Anki.[21]

Those very same form-meaning links are also a dominant feature in one of the most influential basic vocabularies for Latin in Germany, the so-called Bamberg Vocabulary.[22] However, even if we take the high priority of this lexical representation as granted, we are still facing other problems: The Bamberg Vocabulary has been constructed using a closed-source corpus, i.e., the text collection and the formulae applied to it have not been published. In general, we know the authors and works that it was constructed from, but not the exact text passages or text editions. Besides, it is unclear which words were classified as proper names and removed during preprocessing.[23] Therefore, we have too little information to analyze its quality in detail and we have to extrapolate it from vague hints in the corresponding journal paper. In any case, a well-founded estimation of vocabulary size and selection has to be transparent to meet the requirements of serious educational research, especially in times of the generally acknowledged FAIR data principles.[24]

Upon closer examination of the purpose of vocabulary training, research on the didactics of Latin tends to emphasize translation as the core activity in classes.[25] Translation, however, is a highly complex process that involves more than the memorization of form-meaning links: It can be used to identify cultural peculiarities and to make typological comparisons between different languages.[26] For such use cases, simple equations of single words are rather unsuitable. Instead, they can merely act as a preliminary stage for higher-level tasks, e.g., identifying idiomatic multiword expressions,[27] which need special strategies for adequate translation. One important factor in this regard is contextualization, which also facilitates the identification and linguistic analysis of homonymy, polysemy and several other semantic obstacles.[28] This is in line with the basic assumption of distributional semantics, i.e. the context of a word constitutes its meaning.[29] Additionally, it also necessitates low-level morphosyntactic tasks where learners are supposed to highlight specific phenomena in a text (cf. Fig. 1).

[17] González-Fernández / Schmitt (2020), 483.

[18] Daum (2016), 15.

[19] Accessible at https://web.uvic.ca/hrd/latin/wheelock/ (Last access 11.01.2021).

[20] Accessible at http://www.thelatinlibrary.com/101/ (Last access 11.01.2021).

[21] Accessible at https://apps.ankiweb.net/ (Last access 11.01.2021). An example for Latin is the DCC Core Latin collection at https://ankiweb.net/shared/info/180623737 (Last access 11.01.2021).

[22] Utz (2000).

[23] Utz (2000), 152.

[24] Wilkinson et al. (2016).

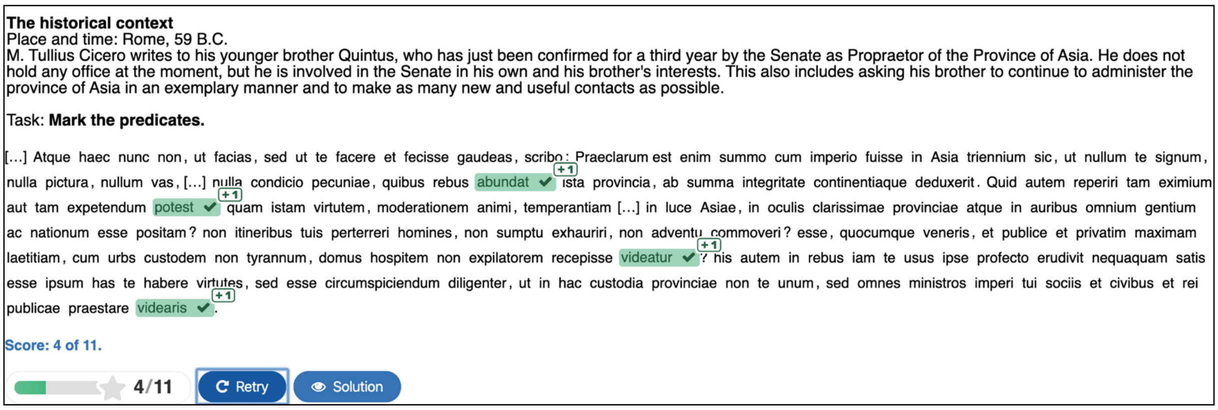
[25] Daum (2016), 76; Große (2015), 191.

[26] Laviosa (2014), 42.

[27] Rayson et al. (2010), 2.

[28] Gardner (2007), 251; Webb (2008), 238; Hagiwara et al. (2009), 556; Helm (2009), 97; Gries / Wulff (2013), 348; Herbelot / Ganesalingam (2013), 443.

[29] Roller et al. (2014), 1025.



The historical context
Place and time: Rome, 59 B.C.
M. Tullius Cicero writes to his younger brother Quintus, who has just been confirmed for a third year by the Senate as Propraetor of the Province of Asia. He does not hold any office at the moment, but he is involved in the Senate in his own and his brother's interests. This also includes asking his brother to continue to administer the province of Asia in an exemplary manner and to make as many new and useful contacts as possible.

Task: **Mark the predicates.**

[…] Atque haec nunc non , ut facias , sed ut te facere et fecisse gaudeas , scribo : Praeclarum est enim summo cum imperio fuisse in Asia triennium sic , ut nullum te signum , nulla pictura , nullum vas , […] nulla condicio pecuniae , quibus rebus abundat ✓ ista provincia , ab summa integritate continentiaque deduxerit . Quid autem reperiri tam eximium aut tam expetendum potest ✓ quam istam virtutem , moderationem animi , temperantiam […] in luce Asiae , in oculis clarissimae provinciae atque in auribus omnium gentium ac nationum esse positam ? non itineribus tuis perterreri homines , non sumptu exhauriri , non adventu commoveri ? esse , quocumque veneris , et publice et privatim maximam laetitiam , cum urbs custodem non tyrannum , domus hospitem non expilatorem recepisse videatur ✓ ? his autem in rebus iam te usus ipse profecto erudivit nequaquam satis esse ipsum has te habere virtutes , sed esse circumspiciendum diligenter , ut in hac custodia provinciae non te unum , sed omnes ministros imperi tui sociis et civibus et rei publicae praestare videaris ✓ .

Score: 4 of 11.

Fig. 1: Text-based exercise for the identification of a morphosyntactic phenomenon, created with H5P.[30]

Designing Digital Exercises for Latin Language Learning

In the following, we will define several quality criteria for exercises, but not in every single detail. Instead, the given example will be explained and analyzed to shed light on the curation process that is linked to the exercise repository. We will particularly highlight the role of contextualization for language acquisition and how this can be implemented in a digital environment. In Fig. 1, students are provided with a heading, a general introduction, a task description, a Latin text, some feedback and control elements. The heading and the introduction are supposed to provide a semantic embedding by describing the historical circumstances of what is portrayed in the Latin text. This is important even for morphosyntactic recognition tasks because previous knowledge of a text's author, topic or background is beneficial for further linguistic analyses.[31] If learners have a basic semantic knowledge (e.g., of the ancient world), they will be able to understand Latin texts more easily,[32] and thus provide a better interpretation or translation. The same correlation also applies to machines when they try to understand natural language.[33] Therefore, the positive influence of prior semantic knowledge on language learning seems to be universal for both humans and machines. This basic interaction is in line with constructivist approaches to learning that emphasize the role of previous knowledge,[34] as well as frame semantics which takes cultural knowledge as the starting point for every language learning process.[35]

The task description should encourage learners to interact with the Latin text. The text is to be presented in a digital format where every word can be clicked on, resulting in its selection for meeting the presented challenge. To assess a learner's performance, the internal digital representation of the text has to be enriched with linguistic annotations, such as part of speech and dependency relation. This way, a learner's selection can be compared to predefined linguistic criteria. In this example, the task is to find predicates, which are defined in Latin grammar books as inflected verb forms that act as the root in the dependency tree of a sentence.[36] However, most of the established linguistic ontologies for describing

[30] Joubel (2018). See also https://h5p.org/ (Last access 11.01.2021), where users can create their own digital interactive exercises using a web application. Such exercises may include advanced features like feedback, hints, learning analytics and various types of media.

[31] Harrison (2010), 9; Mondahl / Razmerita (2014), 341.

[32] Pinkal (1993), 427–428; Fuchs et al. (2015), 211.

[33] Bruni et al. (2014), 38; Punyakanok et al. (2008), 266.

[34] Mvududu / Thiel-Burgess (2012), 110.

[35] Atzler (2011), 61.

[36] Menge et al. (2009), 311.

dependency grammar, such as Universal Dependencies,[37] are not entirely congruent with those used in the Classics.[38] Therefore, a domain-specific language has to be employed to provide a mapping between the two perspectives. In this exercise, the linguistic annotations of part of speech and dependency relation are queried (*Look for a word having the part of speech VERB and serving as root in a dependency tree*) in the background. At the same time, learners are asked to mark all predicates in the text. The comparison of results between both descriptions can be used as an indicator of a learner's performance.

The text's form is determined, to some end, by its corresponding edition. In this case, the Perseus Digital Library was used. The underlying critical text editions do not correspond to the state of the art from a philological perspective, but they conform to the principles of FAIR data and offer a uniform access interface through the Canonical Text Services.[39]

The text's content must be meaningful for learners. In textbooks, this is usually addressed by presenting made-up stories about ancient families.[40] When reading authentic Latin literature, though, the focus is shifted towards passages that cover different curricular standards.[41] Therefore, the text in Fig. 1 deals with the administration of Roman provinces, while still somewhat adhering to the family theme by presenting written communication between two brothers. Besides, it allows learners to improve their form recognition skills in a context-based manner: If they encounter a morphologically ambiguous ending, other words from the same paragraph can be used to eliminate at least some of the possible options.

By building every exercise from ancient literature, i.e., authentic texts written by native speakers, an exercise repository can compensate for the lack of native(-like) language input for learners of historical languages. This introduces the side benefit of highly specialized vocabulary training for single texts, authors or genres. Such a major focus on authentic L2 content in didactic materials is relatively unusual for the teaching of historical languages, which often relies heavily on the learners' L1 for communication in class.[42]

Upon completing an exercise, a status bar together with a twofold numerical representation acts as feedback for the learner's performance. Besides, the given answers are marked visually depending on their quality (green for correct, red for incorrect answers). Solutions that have not been found (i.e., false negatives) remain hidden by default, but can be shown on demand. Retries are possible until the exercise has been completed successfully, but can be disabled entirely. All in all, the quality criteria that we may deduce for the repository can be summarized as follows:

- semantic metadata, e.g., in headings and introductions
- authentic texts created by native speakers
- high quality of digital text editions
- linguistic annotations
- domain-specific language for the interface
- keeping track of learners' actions for the assessment of their language proficiency
- feedback
- control elements, e.g., viewing solutions or retrying old exercises

[37] Nivre et al. (2017).

[38] Kühner / Stegmann (1914); Menge (1914); Menge et al. (2009).

[39] Tiepmar et al. (2014).

[40] Stratenwerth (2012), 264.

[41] Senatsverwaltung für Bildung, Jugend und Sport Berlin (2006), 9–21.

[42] Fuhrmann (2003), 10; Große (2015), 191–202; Kuhlmann (2019), 73.

Furthermore, some studies suggest that interactive exercises provide higher motivation for learners.[43] However, this can be due to the novelty effect of technology-based teaching methods,[44] which are still rather uncommon in Latin pedagogy. Besides, even tasks like in Fig. 1 with their rather basic interaction design must not be underestimated because their interlinked view on vocabulary forces learners to tap into a complex combination of available information. Therefore, they have to be trained to study patterns in language use from various perspectives (cf. Fig. 2). The Keyword-In-Context view offers the possibility to study specific phenomena (e.g., usage of pronouns) in a structured and focused manner, integrating morphological, syntactic and semantic patterns.[45] Such visualizations can quickly become very complex,[46] which is why the example in Fig. 2 is restricted to just a few layers of annotation and contains additional formatting for further clarification (alignment of text passages, colouring, regular geometrical shapes).

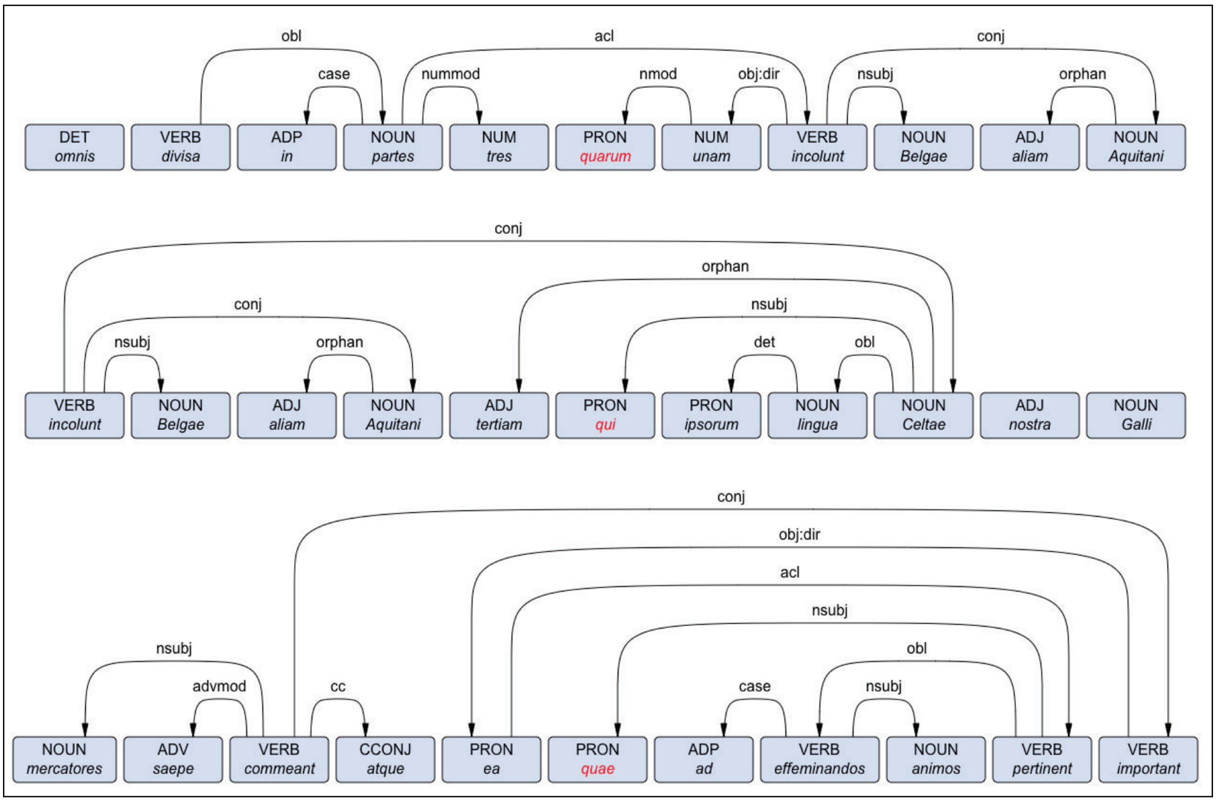


Fig. 2: Keyword-In-Context analysis.[47]

Evaluation

When new lexical knowledge has been acquired, it is notoriously difficult to diagnose that improvement. Traditionally, teachers use lists of word equations, i.e., translations of single words from the foreign to

[43] Harecker / Lehner-Wieternik (2011), 5.

[44] Merchant et al. (2014), 33.

[45] Helm (2009), 97.

[46] Fischl / Scharl (2014), 194.

[47] The visualization was created using CONLLU Viewer (Kleiweg [2020]). It shows parts of speech and dependency relations for three partly overlapping text passages from Caesar's Gallic War, 1.1.1–1.1.3. Each passage is centred around a pronoun, which is highlighted in red. Arrows indicate the direction of a dependency relation from head to tail. The labels refer to the Universal Dependencies tag set. Parts of speech are given for each word in the upper part of a grey box.

the native language.[48] Instead of *native* language, it is more accurate to refer to it as the language in which the *lessons are conducted* because it is usually not the first language for every single learner in terms of acquisition sequence. This indicates a major issue with teaching Latin (at least in Germany): The heavy focus on the German language is usually depicted as beneficial for native speakers[49] as well as second language learners,[50] but it also systematically discriminates against the latter group by relying on the German language for testing purposes. Thus, at least for the diagnosis of lexical knowledge, we should avoid German elements because the separation of comprehension and translation is important for giving differentiated feedback.

In the oral domain (e.g., in teaching modern foreign languages), such abstractions are already present in existing diagnostic tools like the Toolbox Picture Vocabulary Test (TPVT):[51] Participants listen to tape-recorded words and, after each one of them, choose from 4 possible pictures the one that depicts the word's meaning most accurately. In this manner, the language barrier introduced by translation is eliminated, giving way to a more direct estimation of lexical knowledge. Unfortunately, this is hardly applicable to Latin because historical languages are not learned for the purpose of oral communication.[52] A transfer of the TPVT to the Latin domain would therefore need to provide written stimuli. Moreover, since the perceived distance to the target culture (i.e., the Roman Empire) is comparatively large,[53] some concepts may not be easy to depict and convey in an accurate manner. Therefore, such tools are somewhat limited, but still, their basic principles are valuable for designing new evaluations in the teaching of historical languages. These principles also include the internal differentiation of difficulty levels with respect to a learner's current proficiency, in order to avoid floor and ceiling effects.[54] Moreover, individual items in a test should not be weighted equally, but according to their workload and cognitive complexity: An item that asks me to translate a whole text passage will be more challenging than the morphological analysis of a single word. This distinction becomes even more obvious when the general focus of the test is shifted from form-meaning links to reading comprehension,[55] which suggests itself given the particular emphasis on the reception of literature in teaching Latin.

Another problem is the comparatively small learning input between the two tests. In modern language teaching, students can rely on oral practice for accelerated acquisition since it is much faster than written communication, especially in historical languages like Latin.[56] This lack of repetition and intensity has to be compensated by relying on sophisticated educational designs that integrate psychological concepts like the mental lexicon and spreading activation:[57] Thematically related words should be learned (and possibly tested) together. This also implies the rejection of alphabetical word lists for educational purposes.

Finally, in times of blended learning and e-assessment, digital test tools are becoming increasingly popular. They give the impression of objectivity, (social) justice, reliability and efficiency. However, there are many hidden weaknesses in traditional testing that now become obvious in more formalized, compu-

[48] Carter (1997).

[49] Große (2015), 202.

[50] Siebel (2017), 177–178.

[51] Gershon et al. (2013), 54.

[52] Siebel (2017), 18.

[53] Schauer (2019), 182.

[54] Sparrow et al. (2005), 290; Gershon et al. (2013), 56.

[55] Schmitt (2014), 950.

[56] Schirok (2010), 13; Daum (2016), 76.

[57] Bruza et al. (2009), 362/364.

ter-assisted settings. One of them is the lack of a well-defined horizon of expectations for the semantic and morphological parts of translation tasks:[58]

- Which translations (or paraphrases etc.) are appropriate representations of a given target term or concept?
- How do we (consistently) distinguish careless mistakes from a more profound lack of knowledge?
- How do we handle definiteness when translating between language pairs where one part has articles while the other does not?

Recent studies point towards the importance of context and a thorough understanding of its underlying semantics as a prerequisite for adequate translation.[59] This assessment goes beyond the traditional design of vocabulary training, where context was almost entirely eliminated. One approach to reintroduce this complexity in computational settings is a branch of Artificial Intelligence named representation learning, which tries to model each word's semantics by its common textual co-occurrences with other words.[60] However, many special cases are still hard to cover in such frameworks, e.g., multiword expressions.

Feedback

Regardless of whether the evaluation of lexical competence can be automated successfully, we also have to face the challenge of providing high-quality feedback. Usually, this is done in a binary fashion (i.e., correct/incorrect response), with explicit measurements (e.g., a score to be achieved) and a delayed communication of results (e.g., after a few days). For modern languages, there is additional implicit feedback from conversational exercises, e.g., dialogues.[61] Unfortunately, this valuable source of corrective input is mostly unavailable for Latin because of the strong focus on reading. Nevertheless, the same general quality criteria apply: Feedback should be immediate,[62] like a scaffolding,[63] and adapted to a learner's zone of proximal development.[64] This way, students do not just see superficial scores, but a detailed explanation of what they did wrong and what the smallest next step in the right direction might be. Unfortunately, such requirements are hard to meet in both face-to-face and computational settings because they demand a lot of time and/or complex modelling.

A basic, but crucial example consists in the classification of errors: If we do not distinguish between various deviations related to form and meaning,[65] we will fail to give helpful feedback, thus having to fall back to simpler ways of scoring. Besides, a written indication of the locations and types of errors may not be sufficient to encourage corrections. Instead, multimodal feedback (e.g., using videos) may be employed to offer higher incentives for improvement.[66] Furthermore, learners are usually not just interested in their current performance on a single test item, but also on their development over time (cf.

[58] Beatty (2013), 209.

[59] Hummel (2010), 62.

[60] Bengio et al. (2003), 1141; Mikolov et al. (2013), 2; Perez / Cuadros (2017), 51; Wiedemann et al. (2019), 2.

[61] Ellis et al. (2006), 340–341.

[62] Opitz et al. (2011), 7.

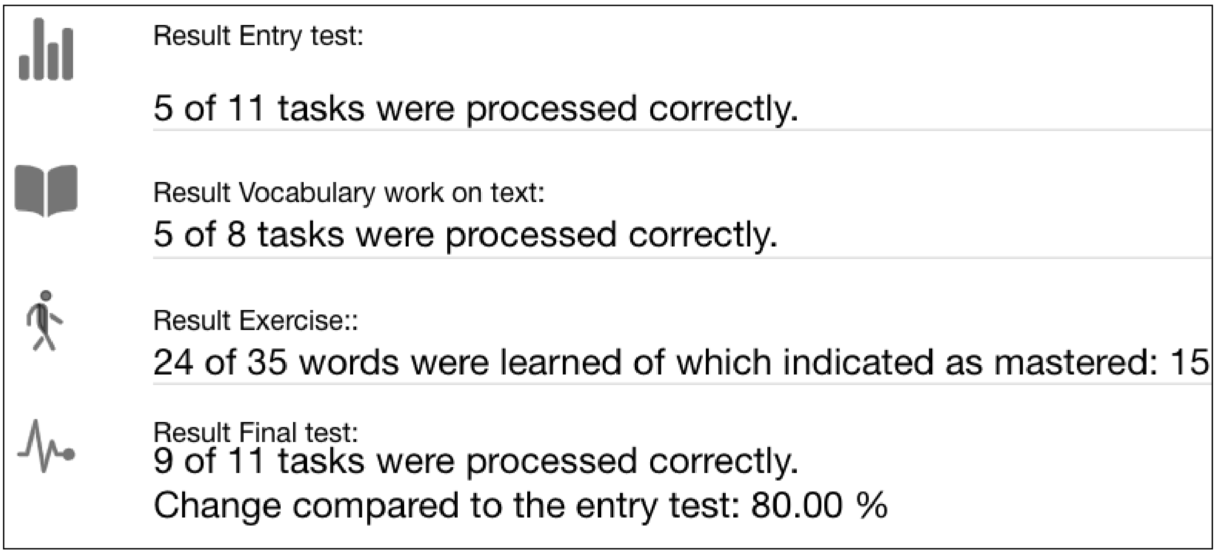
[63] Finn / Metcalfe (2010), 959.

[64] Shabani et al. (2010), 238.

[65] Rudzewitz et al. (2017), 41.

[66] Elola / Oskoz (2016), 71.

Fig. 3).[67] This kind of ipsative assessment helps them keep track of their progress and assume responsibility for their learning.



Result Entry test:

5 of 11 tasks were processed correctly.

Result Vocabulary work on text:

5 of 8 tasks were processed correctly.

Result Exercise::

24 of 35 words were learned of which indicated as mastered: 15

Result Final test:

9 of 11 tasks were processed correctly.

Change compared to the entry test: 80.00 %

Fig. 3: Explicit summative feedback after a structured learning session, for ipsative assessment.[68]

In addition to this explicit kind of individual feedback, consumers of a well-designed vocabulary framework should also have access to criterion-referenced tests. These may include, among others, measurements of textual complexity and overlap with a target vocabulary.[69] This way, learners can quantify and compare their personal level of competence at various points in time. What is more, teachers may choose the most suitable upcoming exercises on an individual basis by browsing the repository of ready-made materials (cf. Fig. 3). Possible use cases for search in this repository include exercises

- of a preferred interaction type, which may correspond to a certain learning style, e.g., only cloze exercises.[70]
- for a specific author, in order to study the author-specific language use.[71]
- for a specific text passage or subcorpus, in order to prepare for a phase of close reading or exams.
- with a low text complexity for the transition from textbook to authentic literature.[72]
- with as few out-of-vocabulary words as possible, which would have to be supplemented with a gloss in exams or for novice learners.[73]

[67] Univio / Pérez (2019), 158.

[68] The four parts belong to a curated 45-minute digital session for Latin vocabulary training (accessible at https://korpling.org/mc/test [Last access 11.01.2021]). The Entry and Final tests were identical, thus enabling the learners to see their progress after a specific intervention (Vocabulary work on text and Exercise). Mastery of words in the Exercise part was indicated through self-assessment by the students using a checkbox.

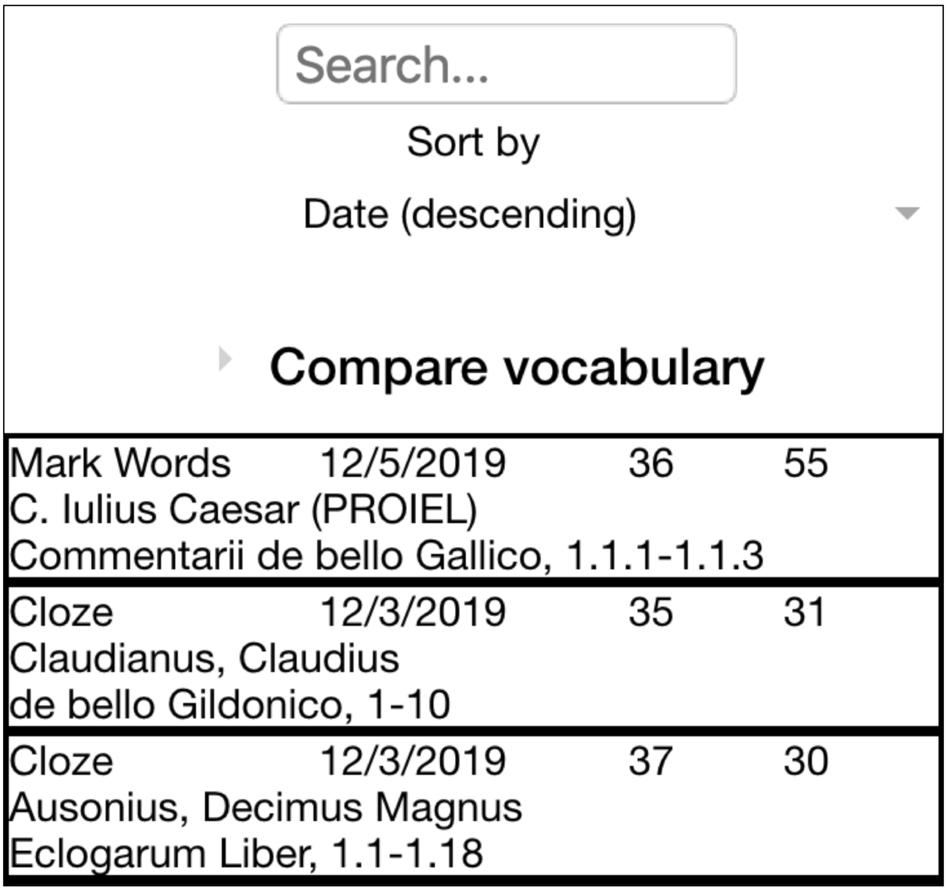
[69] Muccigrosso (2004), 422.

[70] Schmid (2010), 169.

[71] Devine / Stephens (2006), 452; Cordes (2020), 40–41.

[72] Schibel (2013), 115.

[73] Olimpi (2019), 86.



Sort by

Date (descending)

Compare vocabulary

Mark Words C. Iulius Caesar (PROIEL) Commentarii de bello Gallico, 1.1.1-1.1.3	12/5/2019	36	55
Cloze Claudianus, Claudius de bello Gildonico, 1-10	12/3/2019	35	31
Cloze Ausonius, Decimus Magnus Eclogarum Liber, 1.1-1.18	12/3/2019	37	30

Fig. 4: The exercise repository offers various options for filtering: exercise type, date of last access, author, text passage.[74]

Conclusion

Many of the above-mentioned quality criteria for vocabulary learning and exercise repositories have already been considered in our implementation: Almost all vocabulary exercises involve contextualization, i.e., most words are not presented independently, but as part of a collocation, phrase, sentence or even a whole text passage. Depending on their design, multiple items may be combined into a longer sequence, e.g., with increasing levels of difficulty. Indicators of such difficulty are the familiarity with an exercise's vocabulary or the linguistic features of its base text. These are calculated automatically, so users can sort by them and easily compare various materials.

Moreover, the repository makes use of existing high-quality resources such as text editions, annotated corpora and frameworks for interactive digital exercises. Where linguistic information is missing, it tries to add them automatically. This fallback procedure is error-prone, particularly for complex syntactic annotations, thus decreasing the quality of the curation process and the entire repository.

While exercises also include explicit feedback (either immediately after a single exercise or after a longer period of learning), they do so only in a binary fashion. The correct results are shown and teachers may provide a general explanation, but it is not adaptive and thus not suitable to point learners in the right direction. This shortcoming is probably the most important aspect to consider in the development

[74] Every item has measurements of text complexity and the percentage of known words as compared to a reference vocabulary (last two columns).

of future projects. Finally, it seems that the opportunities for individualization offered by a digital learning context seem to ask for an even stronger integration of ipsative assessment. Some of this is already present in the evaluation after the ready-made vocabulary unit, but additional visualizations and a more detailed tracking of learner results are necessary in order to provide a higher overall quality in the curation and reuse of lexical materials.

References

Atzler (2011): J. K. Atzler, Twist in the List: Frame Semantics as Vocabulary Teaching and Learning Tool, PhD Thesis, Austin: University of Texas, 2011, URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/ETD-UT-2011-05-2752/ATZLER-DISSERTATION.pdf?sequence=1&isAllowed=y (Last access 11.01.2021).

Beatty (2013): K. Beatty, Teaching & Researching: Computer-Assisted Language Learning, Routledge, 2013, URL: http://ebook.stkip-pgri-sumbar.ac.id/ebook/bahasa/teaching-researching-computer-assisted-language-learning-pearson-education-esl/download (Last access 11.01.2021).

Bengio et al. (2003): Y. Bengio / R. Ducharme / P. Vincent / C. Jauvin, A Neural Probabilistic Language Model, Journal of machine learning research 3 Feb (2003), 1137–1155.

Beyer / Schulz (2020): A. Beyer / K. Schulz, CALLIDUS – Korpusbasierte, digitale Wortschatzarbeit im Lateinunterricht, in: F. Maier / S. Chronopoulos (eds.), Der Digital Turn in den Altertumswissenschaften, Propylaeum-eBooks, 149–167, 2020, URL: https://doi.org/10.11588/propylaeum.563 (Last access 11.01.2021).

Bruni et al. (2014): E. Bruni / N.-K. Tran / M. Baroni, Multimodal Distributional Semantics, J. Artif. Intell. Res. (JAIR) 49 (2014), 1–47.

Bruza et al. (2009): P. Bruza / K. Kitto / D. Nelson / C. McEvoy, Is There Something Quantum-like about the Human Mental Lexicon?, Journal of Mathematical Psychology 535 (2009), 362–377.

Carter (Aug. 1997): T. G. M. Carter, Latin Vocabulary Acquisition: An Experiment Using Information-Processing Techniques of Chunking and Imagery, English, Dissertation, University of North Texas, Aug. 1997, URL: https://digital.library.unt.edu/ark:/67531/metadc277583/m1/11/ (Last access 11.01.2021).

Cordes (2020): L. Cordes, Wenn Fiktionen Fakten schaffen. Faktuales und fiktionales Erzählen in den spätantiken Panegyrici Latini, Deutsch, in: D. Breitenwischer / H.-M. Häger / J. Menninger (Hrsgg.), Faktuales und fiktionales Erzählen II. Geschichte – Medien – Praktiken, Baden-Baden 2020, 31–56, URL: https://doi.org/10.5771/9783956505126-31 (Last access 11.01.2021).

Crossley et al. (2010): S. A. Crossley / T. Salsbury / D. S. McNamara, The Development of Semantic Relations in Second Language Speakers: A Case for Latent Semantic Analysis, Vigo International Journal of Applied Linguistics 7 (2010), 55–74.

Dascalu et al. (2017): M. A. Dascalu / G. S. Gutu / S. S. Ruseti / I. S. Cristian Paraschiv / P. Dessus / D. A. Mcnamara / S. A. Crossley / S. A. Trausan-Matu, ReaderBench: A Multi-Lingual Framework for Analyzing Text Complexity, in: É. Lavoué / H. Drachsler / K. Verbert / J. Broisin / M. Pérez-Sanagustín (Hrsgg.), Data Driven Approaches in Digital Education, Proc 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings, Tallinn, Estonia 2017 606–609, URL: https://hal.archives-ouvertes.fr/hal-01584870 (Last access 11.01.2021).

Daum (2016): M. Daum, Wortschatz und Lehrbuch: Ein Kriterienkatalog für die Wortschatzkonzeption in Lateinlehrwerken, vol. 2, Ars Didactica. Marburger Beiträge zu Studium und Didaktik der Alten Sprachen, Propyläum-eBooks 2016, URL: https://doi.org/10.11588/propylaeum.609 (Last access 13.01.2021).

Devine / Stephens (2006): A. M. Devine / L. D. Stephens, Latin Word Order: Structured Meaning and Information, Oxford University Press, 2006, URL: https://books.google.de/books?hl=en%5C&lr=%5C&id=WY2Nhc3HY3sC (Last access 11.01.2021).

Ellis et al. (June 2006): R. Ellis / S. Loewen / R. Erlam, Implicit and Explicit Corrective Feedback and the Acquisition of L2 Grammar, Studies in Second Language Acquisition 282 (2006), 339–368.

Elola / Oskoz (2016): I. Elola / A. Oskoz, Supporting Second Language Writing Using Multimodal Feedback, Foreign Language Annals 491 (2016), 58–74.

Finn / Metcalfe (2010): B. Finn / J. Metcalfe, Scaffolding Feedback to Maximize Long-Term Error Correction, Memory & Cognition 387 (2010), 951–961.

Fischl / Scharl (2014): D. Fischl / A. Scharl, Metadata Enriched Visualization of Keywords in Context, in: Proceedings of the 2014 ACM SIGCHI Symposium on Engineering Interactive Computing Systems, 193–196, 2014, URL: https://dl.acm.org/doi/pdf/10.1145/2607023.2611451 (Last access 11.01.2021).

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (2004): Freie und Hansestadt Hamburg, Behörde für Bildung und Sport, Rahmenplan Alte Sprachen: Latein, Griechisch. Bildungsplan achtstufiges Gymnasium Sekundarstufe I, 2004, URL: http://epub.sub.uni-hamburg.de/epub/volltexte/2008/600/pdf/LATGRIE_Gy8.pdf (Last access 11.01.2021).

Fuchs et al. (2015): L. S. Fuchs / D. Fuchs / D. L. Compton / C. L. Hamlett / A. Y. Wang, Is Word-Problem Solving a Form of Text Comprehension?, Scientific Studies of Reading 193 (2015), 204–223.

Fuhrmann (2003): M. Fuhrmann, Bildungsziele im Wandel der Zeiten – und worauf soll es jetzt hinaus? Eine nüchterne Standortbestimmung, auch für Latein und Griechisch, Pegasus-Onlinezeitschrift 32 (2003), 1–11, URL: https://doi.org/10.11588/pegas.2003.2.35716 (Last access 11.01.2021).

Gardner (Apr. 2007): D. Gardner, Validating the Construct of Word in Applied Corpus-Based Vocabulary Research: A Critical Survey, Applied Linguistics 282 (2007), 241–265.

Gershon et al. (2013): R. C. Gershon / J. Slotkin / J. J. Manly / D. L. Blitz / J. L. Beaumont / D. Schnipke / K. Wallner-Allen / R. M. Golinkoff / J. B. Gleason / K. Hirsh-Pasek / M. J. Adams / S. Weintraub, NIH Toolbox Cognition Battery (CB): Measuring Language (Vocabulary Comprehension and Reading Decoding), Monographs of the Society for Research in Child Development 784 (2013), 49–69.

González-Fernández / Schmitt (2020): B. González-Fernández / N. Schmitt, Word Knowledge: Exploring the Relationships and Order of Acquisition of Vocabulary Knowledge Components, Applied Linguistics 41 (2020), 481–505.

Gries / Wulff (2013): S. T. Gries / S. Wulff, The Genitive Alternation in Chinese and German ESL Learners: Towards a Multifactorial Notion of Context in Learner Corpus Research, International Journal of Corpus Linguistics 183 (2013), 327–356.

Große (2015): M. Große, Pons Latinus: Latein als reflexionsbasierte Brückensprache im Rahmen eines sprachsensiblen Lateinunterrichts, in: E. M. F. Ammann / A. Kropp / J. Müller-Lancé (Hrsgg.), Herkunftsbedingte Mehrsprachigkeit im Unterricht der Romanischen Sprachen, vol. 17, Frank & Timme, 185–206, 2015, URL: https://www.peterlang.com/downloadpdf/title/64659 (Last access 18.01.2021).

Hagiwara et al. (2009): M. Hagiwara / Y. Ogawa / K. Toyama, Supervised Synonym Acquisition Using Distributional Features and Syntactic Patterns, Information and Media Technologies 42 (2009), 558–582.

Harecker / Lehner-Wieternik (2011): G. Harecker / A. Lehner-Wieternik, Computer-Based Language Learning with Interactive Web Exercises, ICT for Language Learning (2011), 1–5.

Harrison (2010): R. R. Harrison, Exercises for Developing Prediction Skills in Reading Latin Sentences, Teaching Classical Languages. An Online Journal of the Classical Association of the Middle West and South 21 (2010), 1–30.

Helm (May 2009): F. Helm, Language and Culture in an Online Context: What Can Learner Diaries Tell Us about Intercultural Competence?, Language and Intercultural Communication 92 (2009), 91–104.

Herbelot / Ganesalingam (2013): A. Herbelot / M. Ganesalingam, Measuring Semantic Content in Distributional Vectors, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013, 440–445, URL: https://www.aclweb.org/anthology/P13-2078 (Last access 11.01.2021).

Hummel (2010): K. M. Hummel, Translation and Short-Term L2 Vocabulary Retention: Hindrance or Help?, Language teaching research 141 (2010), 61–74.

Joubel (June 2018): A. Joubel, H5P. Create, Share and Reuse Interactive HTML5 Content in Your Browser, Joubel AS, Tromsø, Norway, URL: https://h5p.org/about-the-project (Last access 11.01.2021), June 2018.

Kleiweg (Sept. 2020): P. Kleiweg, Conllu Viewer. A Web-Based Viewer for Documents in the CoNLL-U Format, Computational Linguistics, University of Groningen, Sept. 2020, URL: https://github.com/rug-compling/conllu-viewer (Last access 11.01.2021).

Kuhlmann (2019): P. Kuhlmann, Sprachausbildung, Aufgabenformate und Übungsdidaktik im Lateinstudium, in: S. Freund / L. Janssen (Hrsgg.), Non ignarus docendi. Impulse zur kohärenten Gestaltung von Fachlichkeit und von Mehrsprachigkeitsdidaktik in der Lateinlehrerbildung, Bad Heilbrunn 2019, 66–78, URL: https://www.pedocs.de/volltexte/2019/16910/pdf/Freund_Janssen_2019_Non_ignarus_docendi.pdf (Last access 11.01.2021).

Kühner / Stegmann (1914): R. Kühner / C. Stegmann, Ausführliche Grammatik der lateinischen Sprache, 2. Teil: Satzlehre, vol. 1, Hannover 1914, URL: https://books.google.de/books?id=PIW7chvuopYC (Last access 11.01.2021).

Laviosa (2014): S. Laviosa, Translation and Language Education: Pedagogic Approaches Explored, London / N.Y. 2014, URL: https://books.google.de/books/?id=cp0QngEACAAJ&redir_esc=y (Last access 13.01.2021).

Menge (1914): H. Menge, Repetitorium der lateinischen Syntax und Stilistik: Ein Lernbuch für Studierende und vorgeschrittene Schüler, zugleich ein praktisches Repertorium für Lehrer, Teile 1–2, Wolfenbüttel 1914, URL: https://books.google.de/books?id=Mr0zAQAAMAAJ (Last access 11.01.2021).

Menge et al. (2009): H. Menge / T. Burkard / M. Schauer, Lehrbuch der lateinischen Syntax und Semantik, Darmstadt 2009.

Merchant et al. (Jan. 2014): Z. Merchant / E. T. Goetz / L. Cifuentes / W. Keeney-Kennicutt / T. J. Davis, Effectiveness of Virtual Reality-Based Instruction on Students' Learning Outcomes in K-12 and Higher Education: A Meta-Analysis, Computers & Education 70 (2014), 29–40.

Mikolov et al. (2013): T. Mikolov / K. Chen / G. Corrado / J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv preprint arXiv:1301.3781 (2013), 1–12, URL: https://arxiv.org/abs/1301.3781 (Last access 11.01.2021).

Mondahl / Razmerita (2014): M. Mondahl / L. Razmerita, Social Media, Collaboration and Social Learning–A Case-Study of Foreign Language Learning. Electronic Journal of E-learning 124 (2014), 339–352.

Muccigrosso (2004): J. D. Muccigrosso, Frequent Vocabulary in Latin Instruction, The Classical World 974 (2004), 409–433.

Mvududu / Thiel-Burgess (2012): N. Mvududu / J. Thiel-Burgess, Constructivism in Practice: The Case for English Language Learners, International Journal of Education 43 (2012), 108.

Nivre et al. (Nov. 2017): J. Nivre et al., Universal Dependencies 2.1: Morphologically and Syntactically Annotated Corpora of Many Languages, Nov. 2017, URL: https://hal.inria.fr/hal-01682188 (Last access 11.01.2021).

Olimpi (2019): A. Olimpi, Legere Discitur Legendo: Extensive Reading in the Latin Classroom, Journal of Classics Teaching 2039 (2019), 83–89.

Opitz et al. (2011): B. Opitz / N. K. Ferdinand / A. Mecklinger, Timing Matters: The Impact of Immediate and Delayed Feedback on Artificial Language Learning, Frontiers in Human Neuroscience 5 (2011), 1–9.

Perez / Cuadros (2017): N. Perez / M. Cuadros, Multilingual Call Framework for Automatic Language Exercise Generation from Free Text, in: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, 49–52, URL: https://www.aclweb.org/anthology/E17-3013.pdf (Last access 18.01.2021).

Pinkal (1993): M. Pinkal, Semantik, in: G. Görz (Hrsg.), Einführung in die Künstliche Intelligenz, Bonn 1993, 425–498.

Punyakanok et al. (2008): V. Punyakanok / D. Roth / W.-T. Yih, The Importance of Syntactic Parsing and Inference in Semantic Role Labeling, Computational Linguistics 342 (2008), 257–287.

Rayson et al. (Apr. 2010): P. Rayson / S. Piao / S. Sharoff / S. Evert / B. V. Moirón, Multiword Expressions: Hard Going or Plain Sailing?, Language Resources and Evaluation 441 (2010), 1–5.

Robillard et al. (2014): M. Robillard / C. Mayer-Crittenden / M. Minor-Corriveau / R. Bélanger, Monolingual and Bilingual Children with and without Primary Language Impairment: Core Vocabulary Comparison, Augmentative and alternative communication 303 (2014), 267–278.

Roller et al. (2014): S. Roller / K. Erk / G. Boleda, Inclusive yet Selective: Supervised Distributional Hypernymy Detection, in: Proceedings of COLING 2014 (The 25th International Conference on Computational Linguistics: Technical Papers), 2014, 1025–1036, URL: https://www.aclweb.org/anthology/C14-1097 (Last access 11.01.2021).

Rudzewitz et al. (May 2017): B. Rudzewitz / R. Ziai / K. De Kuthy / D. Meurers, Developing a Web-Based Workbook for English Supporting the Interaction of Students and Teachers, in: Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, Gothenburg, Sweden May 2017, 36–46, URL: https://www.aclweb.org/anthology/W17-0305 (Last access 11.01.2021).

Schauer (2019): M. Schauer, Klasse Klassik: Latein im Klassenzimmer, in: K. Beuter / A. Hlukhovych / B. Bauer / K. Lindner / S. Vogt (Hrsgg.), Sprache und kulturelle Bildung: Perspektiven für eine reflexive Lehrerinnen- und Lehrerbildung und einen heterogenitätssensiblen Unterricht, vol. 9, Bamberg 2019, URL: https://fis.uni-bamberg.de/bitstream/uniba/46841/1/FLB9BeuterSpracheopusse_A3a.pdf (Last access 11.01.2021).

Schibel (2013): W. Schibel, Zur Aneignung Lateinischer Literatur und Sprache, Forum Classicum (2013), 113–124.

Schirok (2010): E. Schirok, Wortschatzarbeit, in: T. Doepner / M. Keip (Hrsgg.), Interaktive Fachdidaktik Latein, Göttingen 2010, 13–34, URL: https://static.onleihe.de/content/vandenhoeck/20141125/978-3-647-26411-0/v978-3-647-26411-0.pdf (Last access 11.01.2021).

Schmid (2010): E. C. Schmid, Developing Competencies for Using the Interactive Whiteboard to Implement Communicative Language Teaching in the English as a Foreign Language Classroom, Technology, Pedagogy and Education 192 (2010), 159–172.

Schmitt (2014): N. Schmitt, Size and Depth of Vocabulary Knowledge: What the Research Shows, Language Learning 644 (2014), 913–951.

Schulz et al. (2020): K. Schulz / A. Beyer / M. Dreyer / S. Kipf, A Data-Driven Platform for Creating Educational Content in Language Learning, in: Proceedings of the Conference on Digital Curation Technologies (QURATOR 2020 – Conference on Digital Curation Technologies), Berlin 2020, URL: http://ceur-ws.org/Vol-2535/paper_9.pdf (Last access 21.01.2021).

Senatsverwaltung für Bildung, Jugend und Sport Berlin (2006): Senatsverwaltung für Bildung, Jugend und Sport Berlin, Rahmenlehrplan für die gymnasiale Oberstufe. Gymnasien, Gesamtschulen mit Gymnasialer Oberstufe, Berufliche Gymnasien, Kollegs, Abendgymnasien. Latein, URL: https://www.berlin.de/sen/bildung/unterricht/faecher-rahmenlehrplaene/rahmenlehrplaene/mdb-sen-bildung-unterricht-lehrplaene-sek2_latein.pdf (Last access 11.01.2021), 2006.

Shabani et al. (2010): K. Shabani / M. Khatib / S. Ebadi, Vygotsky's Zone of Proximal Development: Instructional Implications and Teachers' Professional Development, English language teaching 34 (2010), 237–248.

Siebel (2011): K. Siebel, Lateinischer Wortschatz als Brücke zur Mehrsprachigkeit? Eine Durchsicht des Aufgabenspektrums aktueller Lateinlehrwerke, Pegasus-Onlinezeitschrift XII (2011), 102–132, URL: https://doi.org/10.11588/pegas.2011.1.35346 (Last access 11.01.2021)

Siebel (2017): K. Siebel, Mehrsprachigkeit und Lateinunterricht: Überlegungen zum lateinischen Lernwortschatz, vol. 4, Göttingen 2017, URL: https://books.google.de/books?hl=de&lr=&id=YOMsDwAAQBAJ (Last access 11.01.2021).

Sparrow et al. (Jan. 2005): S. S. Sparrow / T. M. Newman / S. I. Pfeiffer, 8 – Assessment of Children Who Are Gifted with the WISC-IV, in: A. Prifitera / L. G. Weiss / D. H. Saklofske (Hrsgg.), WISC-IV Clinical Use and Interpretation (Practical Resources for the Mental Health Professional), Jan. 2005, 281–298, URL: http://www.sciencedirect.com/science/article/pii/B9780125649315500098 (Last access 11.01.2021).

Stratenwerth (2012): D. Stratenwerth, Ziemlich grundsätzliche Überlegungen zur Konzeption von Lateinischen Lehrbüchern. Unter besonderer Berücksichtigung der ersten Lehrbuchtexte und ein paar konkrete Beispiele, Forum Classicum 2012 (2012), 264–270.

Tiepmar et al. (2014): J. Tiepmar / C. Teichmann / G. Heyer / M. Berti / G. Crane, A New Implementation for Canonical Text Services, in: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities 2014 (LaTeCH), 1–8, URL: https://www.aclweb.org/anthology/W14-0601 (Last access 11.01.2021).

Univio / Pérez (2019): D. J. Univio / A. d. P. Pérez, Ipsative Assessment of Essay Writing to Foster Reflection and Self-Awareness of Progress, in: E. White / T. Delaney (Hrsg.), Handbook of Research on Assessment Literacy and Teacher-Made Testing in the Language Classroom, Hershey, PA 2019, 157–180, DOI: 10.4018/978-1-5225-6986-2.ch009, URL: https://www.igi-global.com/chapter/ipsative-assessment-of-essay-writing-to-foster-reflection-and-self-awareness-of-progress/217152 (Last access 11.01.2021).

Utz (2000): C. Utz, Mutter Latein und unsere Schüler – Überlegungen zu Umfang und Aufbau des Wortschatzes [BWS], Antike Literatur–Mensch, Sprache, Welt. Dialog Schule und Wissenschaft 34 (2000), 146–172.

Van de Loo (2016): T. Van de Loo, Wortschatzarbeit – Neuere Perspektiven und schulische Praxis, Pegasus-Onlinezeitschrift 16 (2016), 131–151, URL: https://doi.org/10.11588/pegas.2016.0.35254 (Last access 11.01.2021).

Waiblinger (1998): F. P. Waiblinger, Überlegungen zum Konzept des lateinischen Sprachunterrichts. Joachim Gruber zum 60. Geburtstag, Forum Classicum 1998 (1998), 9–19.

Webb (2008): S. Webb, The Effects of Context on Incidental Vocabulary Learning, Reading in a foreign Language 202 (2008), 232–245.

Wiedemann et al. (Oct. 2019): G. Wiedemann / S. Remus / A. Chawla / C. Biemann, Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings, arXiv:1909.10430 [cs] (2019), 1–10, URL: https://arxiv.org/abs/1909.10430 (Last access 11.01.2021).

Wilkinson et al. (Mar. 2016): M. D. Wilkinson / M. Dumontier / I. J. Aalbersberg / G. Appleton / M. Axton / A. Baak / N. Blomberg / J.-W. Boiten / L. B. da Silva Santos / P. E. Bourne / J. Bouwman / A. J. Brookes / T. Clark / M. Crosas / I. Dillo / O. Dumon / S. Edmunds / C. T. Evelo / R. Finkers / A. Gonzalez-Beltran / A. J. G. Gray / P. Groth / C. Goble / J. S. Grethe / J. Heringa / P. A. C. ’t Hoen / R. Hooft / T. Kuhn / R. Kok / J. Kok / S. J. Lusher / M. E. Martone / A. Mons / A. L. Packer / B. Persson / P. Rocca-Serra / M. Roos / R. van Schaik / S.-A. Sansone / E. Schultes / T. Sengstag / T. Slater / G. Strawn / M. A. Swertz / M. Thompson / J. van der Lei / E. van Mulligen / J. Velterop / A. Waagmeester / P. Wittenburg / K. Wolstencroft / J. Zhao / B. Mons, The FAIR Guiding Principles for Scientific Data Management and Stewardship, Scientific Data 3 (2016), 1–9.

Author contact information[75]

Konstantin Schulz

Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Tel: 030/2093 9720

E-Mail: schulzkx@hu-berlin.de

[75] The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY 4.0.

Digital Humanities auf dem Weg zu einer Wissenschaftsmethodik: Transparenz und Fehlerkultur*

Charlotte Schubert

Abstract: The methods of the Digital Humanities have been subject to massive criticism for quite some time. There are two main accusations that are made again and again: On the one hand, the Digital Humanities do not lead to new results, but present the familiar in a different guise. On the other hand, the methods of the Digital Humanities even lead to false results. Furthermore, it is concluded that the reproducibility/replication and thus the scientific soundness of the results is questionable. This article deals with this issue of (assumed or actual) incorrectness by analyzing this accusation and presents a proposal for a critical approach to errors that can secure the Digital Humanities their place in scientific methodology.

I. Keine neuen Ergebnisse, aber dafür Fehlerhaftigkeit?[1]

An sich sollte die Definition der Digitalen Geisteswissenschaften, wie sie etwa Sybille Krämer ausgehend von dem Begriff der ‚Verdatung' jüngst zusammengefaßt hat, für sich sprechen können:[2] Krämer hat den Neuigkeitswert von Erkenntnissen in den Digital Humanities daran gemessen, daß Erkenntnisse und Einsichten zu gewinnen sind, die mit nichtdigitalen Methoden entweder ganz schwierig oder überhaupt nicht zu erreichen sind.[3] Der Vorwurf gegenüber den Digital Humanities ist aber nun, sie führten nicht zu neuen Ergebnissen, sondern würden lediglich Bekanntes in anderem Gewand präsentieren.[4] Dieser Vorwurf läßt sich jedoch leicht entkräften. Denn ob ein Ergebnis neu oder altbekannt ist, ist unter Fachkollegen recht schnell zu klären. Die Neuigkeitsanforderung, die die Kritiker an die Digital Humanities richten, hat sich längst als erfüllbar erwiesen. In diversen Projekten und Publikationen sind dazu

[*] Das hier behandelte Thema geht auf einen Vortrag zurück, den ich am 21.11.2019 im Rahmen des Symposiums „Wozu Digitale Geisteswissenschaften. Innovationen, Revisionen, Binnenkonflikte" (DFG-geförderte Symposienreihe „Digitalität in den Geisteswissenschaften") gehalten habe. Das Symposium fand an der Leuphana Universität Lüneburg unter der Leitung von Sybille Krämer, Claus Pias und Martin Huber statt. An dieser Stelle möchte ich Sybille Krämer für ihre vielen Anregungen und weiterführenden Hinweise zu dem Thema sehr herzlich danken.

[1] Thomas (2016), 525: „Yet, paradoxically, the 20-year surge in the digital humanities – from 1993 to 2013 – has produced relatively little interpretive or argumentative scholarship. In this first phase of the digital humanities, scholars produced innovative and sophisticated hybrid works of scholarship, blending archives, tools, commentaries, data collections, and visualizations. For the most part in the disciplines, however, few of these works have been reviewed or critiqued. Because the disciplines expect interpretation, argument, and criticism, it could be argued that digital humanists have not produced enough digital interpretive scholarship, and what we have produced has not been absorbed into the scholarly disciplines." Vgl. ebf. van Zundert (2016), 331–347.

[2] Krämer (2018), 5–11, online: https://doi.org/10.11588/dco.2017.0.48490 (17.7.2021).

[3] Krämer (2018), 6.

[4] Da (2019a) in https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986 (17.7.2021). Vgl. Da (2019b), 601–639 und mit einer Übersicht zu den Reaktionen auf den Beitrag von Nan Z. Da vgl. Schubert (2019a) https://doi.org/10.11588/dco.2019.2.72004 (17.7.2021).

Ergebnisse vorgelegt worden, die von der Autorschaftszuweisung[5] über Text- und Diskursanalysen[6] bis hin zu Simulationsexperimenten reichen.[7]

Die Neuigkeitsanforderung hat jedoch noch eine zweite Seite. Ein in den Naturwissenschaften gängiges methodisches Vorgehen, die Bestätigung eines Ergebnisses mit einer zweiten, methodisch anderen, von der ersten unabhängigen Methode, ist in den Geisteswissenschaften bisher nicht üblich. Die Bestätigung eines bereits bekannten Ergebnisses, das mit den Methoden der klassischen, historisch-kritischen Analyse erzielt wurde, mit einer anderen Methode, nämlich aus dem Digital-Humanities-Bereich, die Erkenntnisse auf einem unabhängig von der klassisch-hermeneutischen Methode operierenden Weg bestätigen kann, bietet den Geisteswissenschaften eine völlig neue Perspektive. Sie kann den Geisteswissenschaften einen Weg eröffnen, über das Belegen zu einer dem Beweisen im naturwissenschaftlichen Sinn gleichartigen und intersubjektiv stärkeren Bestätigungspraxis zu kommen. Ein solches Vorgehen – das Bestätigen eines bekannten Ergebnisses durch Methoden der Digital Humanities –, führt nicht zu neuen Erkenntnissen. Es handelt sich jedoch um einen Proof of Concept für die Digital Humanities und deren Beitrag zu den Geisteswissenschaften. Die Bedeutung dieses Vorgehens spielt derzeit eine größere Rolle in den Digital Humanities als in den mit klassischen Methoden vorgehenden geisteswissenschaftlichen Fächern. In den momentanen fachlichen Diskussionen über Sinn und Nutzen der Digital Humanities auf der Seite der Geisteswissenschaften wird die Möglichkeit, die Bestätigungsfunktion des Proof of Concept durch die Digital Humanities zu nutzen, noch nicht oft genutzt. Um diesem Verfahren den Stellenwert zuzuweisen, der ihm einen methodischen Rang als eigene Methode verleiht, sind einige anspruchsvolle Voraussetzungen zu erfüllen:

- Ein Proof of Concept, der durch die Entwicklung und/oder Anwendung der Methoden der Digital Humanities erzielt worden ist, muß sich auf fachlich relevante Ergebnisse oder Diskussionen beziehen.

- Ein Proof of Concept muß auf einem methodisch reflektierten Konzept beruhen, das transparent und nachvollziehbar ist.

- Ein Proof of Concept muß skalierbar sein, so daß die spezifischen, quantitativen Ausweitungsmöglichkeiten, die die Transformation in die ‚Verdatung' mit sich bringt, nutzbar gemacht werden.[8]

Unter diesen Voraussetzungen kann der Proof of Concept für beide Seiten ein Gewinn werden: Für die Geisteswissenschaften als eine neue und erweiterte Bestätigungspraxis und für die Digital Humanities als Anerkennung ihres methodischen Beitrags in den Geisteswissenschaften. Dies ist keineswegs Zukunftsmusik, sondern bereits in der wissenschaftlichen Praxis erprobt und mit durch die Fachcommunity akzeptierten Beiträgen zu belegen. Insbesondere die klassischen Altertumswissenschaften, die auf 2500 Jahre Erfahrung in Wandel und Möglichkeiten des kritischen Umgangs mit Texten zurückblicken können, sind prädestiniert, um hier methodisch beispielgebend zu wirken. *Pars pro toto* sei dazu auf die von Werner Riess und seinen Mitarbeitern publizierten Beiträge zu den Machbarkeitsstudien für ERIS verwiesen.[9]

[5] Der spektakulärste Fall einer erfolgreichen Autorschaftszuweisung ist die Identifizierung des Autors Robert Galbraith als J.K. Rowling durch Juola (2015), 101–113.

[6] Z.B. Schubert / Weiß (2015), 447–471 und Schubert (2018), 79–92.

[7] Z.B. Warnking (2015) und ders. (2016), 45–90; Schäfer (2019), 22–33, online: https://doi.org/10.11588/dco.2019.1.60564 (17.7.2021).

[8] Riess (2019), 4–27 und online: https://doi.org/10.11588/dco.2019.2.72018 (17.7.2021) ist ein überzeugendes Beispiel für ein Proof of Concept. Vgl. auch Riess 2020 (wie in Anm. 9).

[9] Riess (2020), 445–473; Diemke (2020), 57–74 und online: https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/77663/71565 (17.7.2021).

Der zweite Vorwurf, die Methoden der Digital Humanities führten zu falschen bzw. nicht validen Ergebnissen, ist der schwerwiegendere von den beiden Vorwürfen, mit denen sich die Digital Humanities auseinandersetzen müssen. Dieser Vorwurf ist nicht ganz neu, aber er ist in den letzten Jahren mit neuer und erheblicher Schärfe in der Debatte erhoben worden. Die neueren Beiträge, vor allem der Artikel von Nan Z. Da vom 27.3.2019 in ‚The Chronicle of Higher Education" unter der Überschrift „The Digital Humanities Debacle. Computational methods repeatedly come up short" ist, wie das Echo sehr deutlich zeigt, als ein Frontalangriff auf die Digital Humanities aufgefaßt worden.[10] Der Vorwurf lautet, eine spezifische Fehlerhaftigkeit der Digital Humanities sei deren Charakteristikum.[11] Aber nicht nur die Qualität der Daten, die mittlerweile als Problem- und Arbeitsfeld ausreichend thematisiert worden ist, sondern vor allem die Arbeitsweisen, mit der die zugrunde gelegten Daten erschlossen werden, lassen sehr schnell erkennen, wo das Problem liegt. Dies soll hier anhand zweier Praktiken – der Verwendung von Stoppwortlisten und von Metadaten – *pars pro toto* erläutert werden.

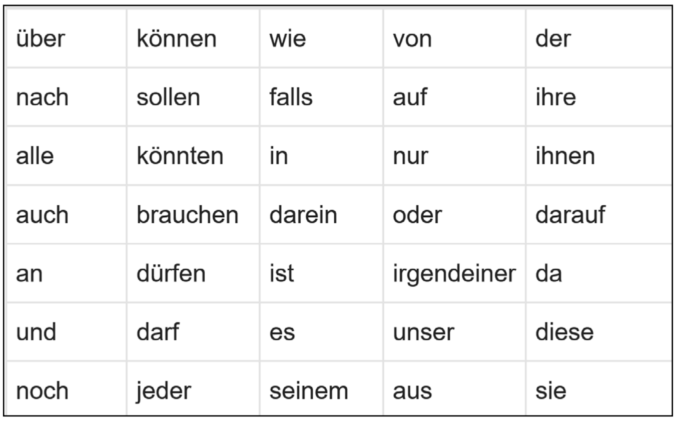
Sowohl die Entwicklung als auch die kompetente Anwendung einer algorithmen-gestützten Auswertung eines oder mehrerer Texte setzt eine Reflexion auf den Zusammenhang zwischen Wortlaut und Bedeutung sprachlicher Äußerungen voraus. Diese, für Textwissenschaftler selbstverständliche, Voraussetzung wird allerdings in vielen Bereichen der mit Texten arbeitenden Digital Humanities mißachtet.

Die Verwendung von Stoppwörtern

Als Stoppwörter werden Wörter bezeichnet, die bei einer Volltextindexierung nicht beachtet werden, da sie sehr häufig auftreten und ihnen für gewöhnlich keine Relevanz für die Erfassung des Dokumenteninhalts zugebilligt wird. Die Stoppwörter befinden sich in der Regel auf einer festen oder berechneten Liste, werden bei der Textverarbeitung aus dem Text entfernt und für die Textanalyse nicht indexiert. Allen Stoppwörtern ist gemeinsam, daß sie vor allem grammatikalische/syntaktische Funktionen übernehmen und daher – so die gängige Meinung – keine Rückschlüsse auf den Inhalt des Dokumentes zulassen. Eine weitere Gemeinsamkeit ist ihre Häufigkeit: Sie treten in jedem Dokument sehr oft auf und kommen in sehr vielen Dokumenten vor, wodurch sie, wenn man sie einbezöge, bei der Erschließung der Dokumente einen hohen Aufwand verursachen würden.

[10] Da (2019a): https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986 (17.7.2021). Vgl. Da. (2019b), 601–639 und Kirby (2019): https://doi.org/10.29173/iq926 (17.7.2021) sowie Rizvi (2018), 401–418 und online: http://doi.org/10.1093/llc/fqy038 (17.7.2021). Dazu Schubert (2019a), 1–3 und online: https://doi.org/10.11588/dco.2019.2.72004 (17.7.2021).

[11] Da (2019a) in https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986 (17.7.2021): „Not only has this branch of the digital humanities generated bad literary criticism, but it tends to lack quantitative rigor. Its findings are either banal or, if interesting, not statistically robust. The problem appears to be structural. In order to produce nuanced and sophisticated literary criticism, CLS [sc. Computational Literary Studies, C.S.] must interpret statistical analysis against its true purpose; conversely, to stay true to the capacities of quantitative analysis, practitioners of CLS must treat literary data in vastly reductive ways, ignoring everything we know about interpretation, culture, and history. Literary objects are too few, and too complex, to respond interestingly to computational interpretation — not mathematically complex, but complex with respect to meaning, which is in turn activated by the quality of thought, experience, and writing that attends it." Eine Zusammenstellung verschiedener Antworten auf Nan Z. Da findet sich hier: https://demo.hedgedoc.org/s/rJ_YoK_cH (17.7.2021).



über	können	wie	von	der
nach	sollen	falls	auf	ihre
alle	könnten	in	nur	ihnen
auch	brauchen	darein	oder	darauf
an	dürfen	ist	irgendeiner	da
und	darf	es	unser	diese
noch	jeder	seinem	aus	sie

Abb. 1: Beispiel einer Stoppwortliste (aus: https://wiki.infowiss.net/Stoppwort, abgerufen am 17.7.2021).

Das Beispiel zeigt einen Ausschnitt aus einer Stoppwortliste für das Deutsche mit einem Teil der üblicherweise als Stoppwörter betrachteten Wörter. Aber nun sind gerade Wort wie „sein“, „sollen“, „dürfen“ bspw. im philosophischen Bereich sinntragende Wörter. So setzt sich etwa Platon in dem Dialog *Sophistes* mit der Frage auseinander, ob man das Nicht-Seiende – da es bekanntlich nicht ist – sagen kann.[12] Folgende Formulierung ist eine zentrale Passage in diesem Dialog:

τὸ μὴ ὂν εἶναι (Plat. *Sophistes* 237 a3f.)
„daß das Nicht-Seiende sei“

Diese Formulierung besteht nun aus lauter Stoppwörtern (Artikel, Negationspartikel, Kopula) und ist daher mit den üblichen Verfahren, die Stoppwortlisten verwenden, gar nicht auffindbar! Das Problem ist hier die Annahme, daß die Stoppworte wie die Artikel und Partikel, auch εἶναι in der geläufigen Verwendung als Kopula, eine rein grammatische Funktion besitzen und nichts zum Sinn beitragen.

Das Problem ist lösbar: Wenn man Platons Werk kennt, dann ist auch die Bedeutung der Kopula εἶναι geläufig. Daher muß – und zwar von Beginn einer Konzeptionierung, Modellierung oder Analyse an – darauf geachtet werden, daß man für entsprechende Textgattungen eben nicht mit automatisch erzeugten Stoppwortlisten arbeitet, sondern sie entweder fach- und sachgerecht erarbeitet oder vollständig beiseite läßt. Die in der Verwendung von Stoppwortlisten implizierten Fehler wären sehr leicht durch an- und abschaltbare bzw. auch anpaßbare Stoppwortlisten zu beseitigen, eine Praxis, die jedoch selten bis gar nicht eingesetzt wird.[13]

Die Verwendung von Metadaten

Neben der Verwendung von Stoppwörtern ist eine weitere, nicht nur allgemein verbreitete, sondern mittlerweile institutionalisierte Praxis, der Gebrauch und die Auswertung von Metadaten.[14] Im Bereich der bibliographischen Metadaten und insbesondere im Data Profiling in der Informatik werden Metadaten als Klassifikationen eher unhinterfragt verwendet. Insbesondere das Data Profiling ist ein Prozeß, der

[12] Das nachfolgende Beispiel ist entnommen aus dem Beitrag von Rautenberg (2019), 111–123, online: https://doi.org/10.11588/propylaeum.451 (17.7.2021).

[13] Vgl. dazu die Webportale mit an- und abschaltbaren Stoppwortlisten zu Digital Plato https://digital-plato.org/ und www.eaqua.net (17.7.2021). In der Paraphrasensuche von Digital Plato ist die Stoppwortliste dynamisch anpaßbar.

[14] Das nachfolgende Beispiel ist eine gekürzte und aktualisierte Version von Schubert (2019b), 4–21 und online: https://doi.org/10.11588/dco.2019.1.59356 (17.7.2021).

Metadaten aus Datenbanken analysiert, um deren Metadaten gegebenenfalls zu korrigieren, der jedoch automatisch durchgeführt wird.

Grundsätzlich böte das Arbeiten mit Textdaten die Chance, verschiedene Datentypen wie die Daten historischer Ereignisse, Werkzuordnungen, geographische Informationen u.v.a.m. auf wissenschaftlicher Basis zu klassifizieren. Die Komplexität von Textdaten ist jedoch oft mit Unsicherheiten wie etwa unklaren Lebensdaten und -orten oder umstrittenen Werkidentifikationen verbunden. Daraus ergibt sich, daß die einzelnen Datentypen (chronologische Einordnung, geographische Verteilungen, Werkzuordnungen, lexikographische Anordnung) wiederum eine Vielzahl von Einzelaspekten umfassen können, d.h. daß diese Unsicherheiten bei der Aufbereitung von Textdaten einbezogen werden müßten. Dies wäre durchaus möglich, allerdings läßt die heute übliche Praxis der Metadatenauszeichnung nicht erkennen, daß die unterschiedlichen wissenschaftssystematischen und wissenschaftshistorisch-methodischen Positionen in der Praxis berücksichtigt werden.

Die diversen Interoperabilitätsbemühungen im Hinblick auf die unterschiedlichen Metadatenstandards und -ontologien verweisen an sich schon auf das Dilemma:[15] Die Gemeinsame Normdatei (GND) der Deutschen Nationalbibliothek, die von allen deutschen Bibliotheken und insbesondere zur Vernetzung von deren Informationsressourcen verwendet wird, führt daher für die Deutsche Nationalbibliothek, alle deutschsprachigen Bibliotheksverbünde mit den angeschlossenen Bibliotheken, die Zeitschriftendatenbank (ZDB) und zahlreiche weitere Einrichtungen die Metadaten für Personen, Körperschaften, Konferenzen, Geographika, Sachschlagwörter und Werktitel, die vor allem zur Katalogisierung von Literatur in Bibliotheken dienen.[16]

Allerdings tritt hier eine Kontingenz zutage, die von denen der lexikographischen Ordnungsverfahren her bekannt sein müßte: Die Voraussetzungen von Etikettierungen oder Labels oder Indexeinträgen sind immer begriffsbezogene Datenkategorien, die aus fachspezifischen Kontexten stammen, die durch automatisierte Extraktion nicht zuverlässig erfaßt werden können. Gleichwohl wird der Anspruch erhoben, daß bibliographische Metadaten von den fachspezifischen Voraussetzungen und deren Kontingenzen unabhängig seien, da man bibliographisch-administrative Daten einerseits und inhaltsbeschreibende bzw. fachliche Daten andererseits unterscheiden könnte. Erstere gäben Informationen zur Verwaltung der Daten aus den Publikationen selbst, letztere hingegen beschrieben einzelne Aspekte oder Datensätze genauer, und inhaltsbeschreibende Metadaten seien grundsätzlich – disziplinspezifisch – unterschiedlich aufgebaut.

Welche Fallen jedoch trotz dieser an sich nicht falschen Unterscheidung der Metadaten in dem Verfahren lauern, ist bereits früher in dieser Zeitschrift ausführlich dargestellt worden.[17] Hier soll dies daher nur kurz rekapituliert und das damalige Ergebnis mit einer weiteren Methode der digitalen Textanalyse untermauert werden.

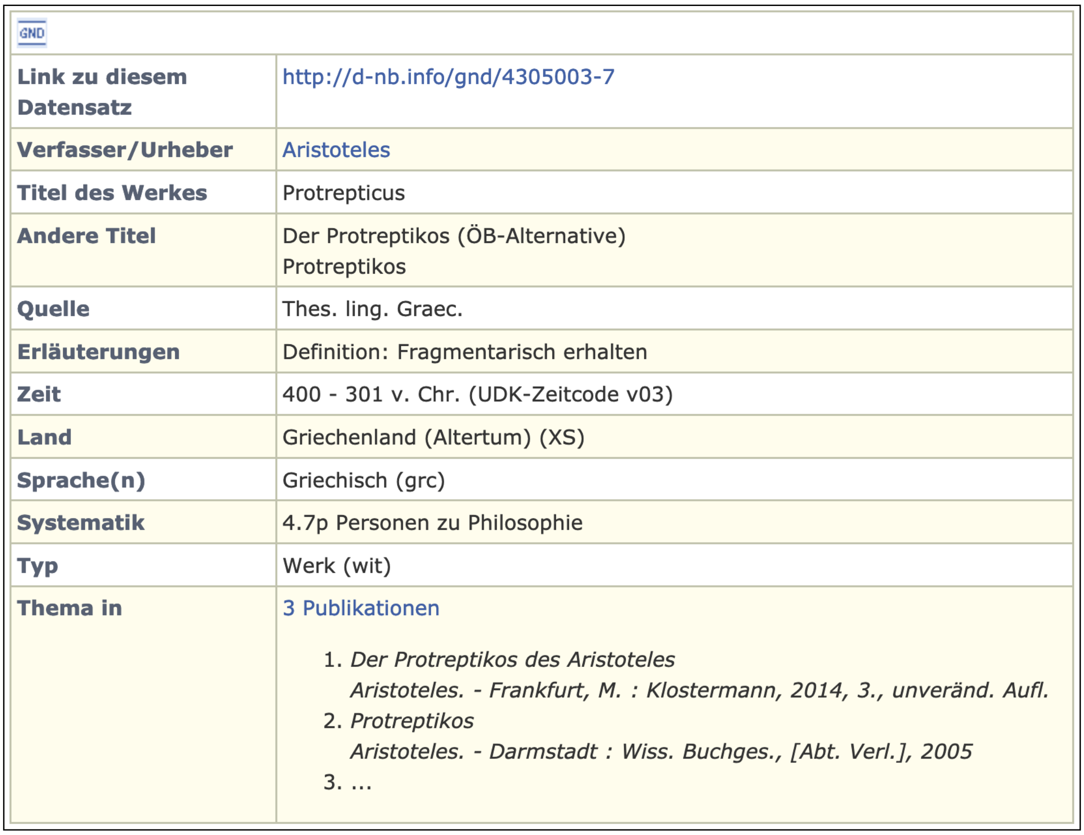
Von dem Neuplatoniker Iamblich von Chalkis (Mitte des 3. Jh. n. Chr. – ca. 320/325 n. Chr.), ist ein zehnbändiges Werk über die Lehre des Pythagoras erhalten. Dessen zweiter Band ist ein sog. *Protreptikos*, eine Mahnschrift in Form des Aufrufes, Philosophie zu betreiben. Diese Gattung war in der Antike sehr beliebt und stand zur Zeit des Iamblich bereits in einer langen Tradition berühmter Vorgänger. Insofern überrascht es nicht, in Iamblichs *Protreptikos* zahlreiche Zitate aus Werken anderer

[15] Die Angaben zu den verwendeten Metadatenformaten und den internationalen Interoperabilitätsbemühungen der Deutschen Nationalbibliothek: https://www.dnb.de/DE/Professionell/Standardisierung/Standards/standards_node.html#%5BAnkerMARC21%5D (17.7.2021).

[16] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html (17.7.2021).

[17] Die Ergebnisse zu Iamblichs *Protreptikos* sind in Schubert (2017), 17–48 publiziert, online: https://books.ub.uni-heidelberg.de/propylaeum/catalog/book/257 (17.7.2021).

Autoren zu finden. Aus diesem Befund ist die These entwickelt worden, daß sich in dem iamblichischen *Protreptikos* auch Teile des verlorenen, aber in der Antike außerordentlich berühmten aristotelischen *Protreptikos* finden ließen. Daher haben verschiedene Altphilologen den Versuch unternommen, aus dem iamblichischen *Protreptikos* den aristotelischen *Protreptikos* zu rekonstruieren und verschiedenste Editionen dieses – eigentlich verlorenen, also nichtexistenten – aristotelischen *Protreptikos* publiziert. Diese Editionen sind größtenteils aus Textpassagen des iamblichischen *Protreptikos* zusammengesetzt und werden in den Bibliothekskatalogen unter dem Namen des Autors Aristoteles rubriziert. So finden sie sich sowohl in den digitalen Textdatenbanken (z.B. in der altgriechischen Textdatenbank TLG[18]) wie auch in den Metadaten-Einträgen der Bibliotheksdatenbanken. Der entsprechende Eintrag der GND führt diesen „Protreptikos des Aristoteles“ folgendermaßen auf:



GND	
Link zu diesem Datensatz	http://d-nb.info/gnd/4305003-7
Verfasser/Urheber	Aristoteles
Titel des Werkes	Protrepticus
Andere Titel	Der Protreptikos (ÖB-Alternative) Protreptikos
Quelle	Thes. ling. Graec.
Erläuterungen	Definition: Fragmentarisch erhalten
Zeit	400 - 301 v. Chr. (UDK-Zeitcode v03)
Land	Griechenland (Altertum) (XS)
Sprache(n)	Griechisch (grc)
Systematik	4.7p Personen zu Philosophie
Typ	Werk (wit)
Thema in	3 Publikationen 1. *Der Protreptikos des Aristoteles* *Aristoteles. - Frankfurt, M. : Klostermann, 2014, 3., unveränd. Aufl.* 2. *Protreptikos* *Aristoteles. - Darmstadt : Wiss. Buchges., [Abt. Verl.], 2005* 3. ...

Abb. 2: http://d-nb.info/gnd/4305003-7 (19.7.2021).

Die früher bereits publizierte Visualisierung auf der Basis der Metadaten und der N-Gramm-Analyse hat gezeigt,[19] wie die Metadaten als Kernstück der bibliographischen Klassifizierung zu falschen Ergebnissen führen müssen, da aufgrund der bibliographischen Angaben natürlich die Aristoteles-Zuweisung hervorsticht.[20] Ohne Zugrundelegung der vorgegebenen Autorzuweisung für den iamblichischen *Protreptikos* bzw. ohne Einbeziehung der rekonstruierten Editionen ergibt sich jedoch, daß der Text des Iamblich keinerlei textuelle Verbindung zu dem Werk des Aristoteles hat, jedoch sehr maßgebliche zu

[18] http://stephanus.tlg.uci.edu (17.7.2021).

[19] Vgl. Schubert (2019b), 4–21 und online: https://doi.org/10.11588/dco.2019.1.59356 (17.7.2021).
Zugrunde liegt die Extraktion der Parallelstellensuche aus eAQUA (N-Gramm-Analyse auf der Basis von fünf exakt gleichen Wort-N-Grammen), deren Metadaten mit Hilfe des Netzwerkvisualisierungsprogramms Gephi (https://gephi.org/ [17.7.2021]) visualisiert wurden. Die verwendeten Metadaten sind Autor- und Werkname – zwei Metadatenkategorien, die nicht nur das grundlegende Gerüst aller Metadatenmodelle sind, die sich auf Textdaten beziehen, sondern heute praktisch immer unhinterfragt eingesetzt werden.

[20] Dazu Abb. 4 in Schubert (2019b), 4–21 und online: https://doi.org/10.11588/dco.2019.1.59356 (17.7.2021).

demjenigen Platons.[21] Auch eine Überprüfung des früheren Ergebnisses mit einer anderen Methode, der Buchstaben-Trigramm-Analyse aus der Stilometrie, bestätigt dieses Ergebnis.[22]

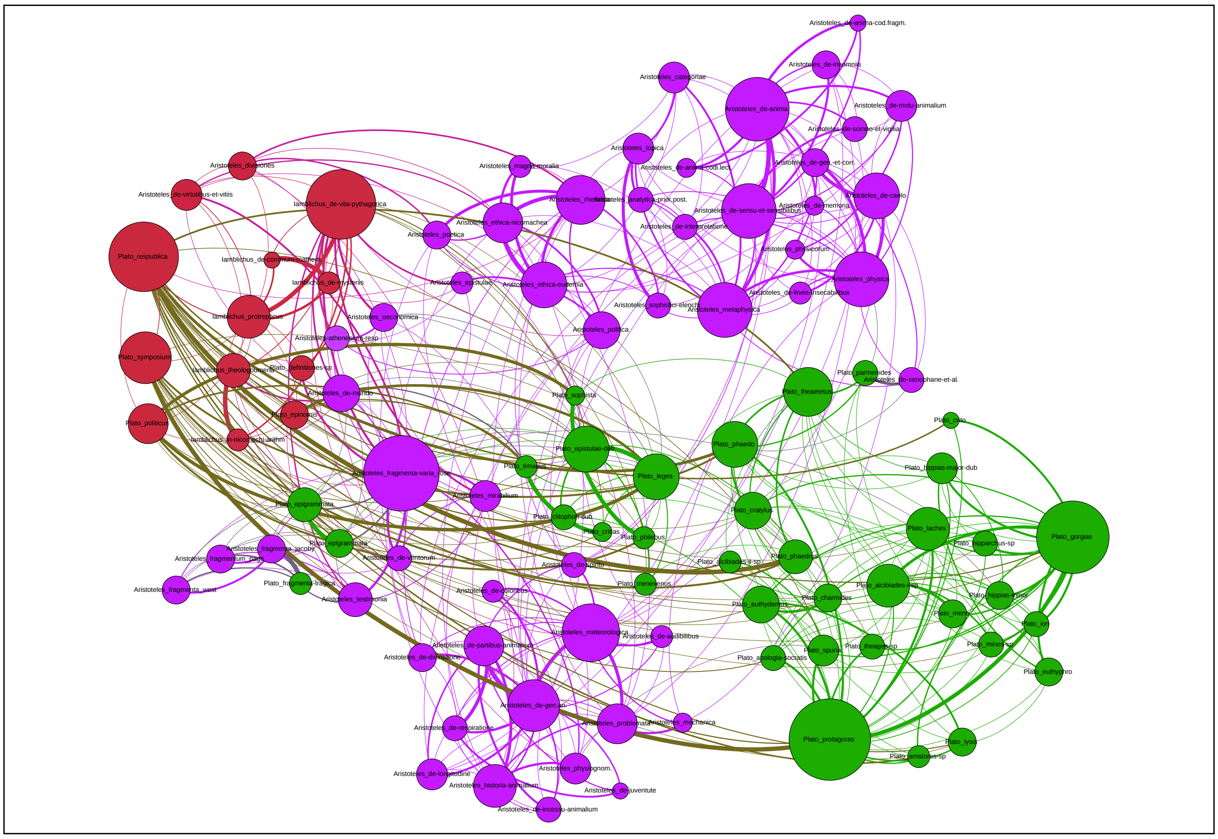


Abb. 3: Visualisierung der Buchstaben-Trigramm-Analyse aus StyloAH in Gephi; rot markiert ist das Netzwerk aus nächsten Nachbarn des Knoten Iamblichs *Protreptikos* mit den entsprechenden Verbindungen zwischen den Knoten.

Die Visualisierung der mit StyloAH durchgeführten Untersuchung eines Textkorpus mit Gephi,[23] das die Werke des Iamblich, Platons und Aristoteles (ohne die moderne Rekonstruktionsedition des *Protreptikos*) in einem Textkorpus zusammenstellt, um Verbindungen zu erkennen, zeigt die Nähe des iamblichischen *Protreptikos* zu Platon, insbesondere zu der platonischen *Politeia*, während für Aristoteles lediglich eine Verbindung zwischen dem iamblichischen *Protreptikos* und der heute als unecht klassifizierten aristotelischen Schrift *De virtutibus et vitiis* sowie der pseudo-aristotelischen Sammlung der *Divisiones* angezeigt wird.[24]

Dieses Ergebnis verweist auf zwei Aspekte: Zum einen wird das frühere Ergebnis der kombinierten Metadaten-/Wort-N-Grammanalyse durch eine buchstabenbasierte N-Grammanalyse erneut bestätigt und erhöht somit die Plausibilität mit Bezug auf die oben genannte Forderung nach dem Einsatz von zwei (oder sogar mehr) unabhängig voneinander operierenden Methoden. Zum anderen kann die Un-

[21] Dazu Abb. 6a in Schubert (2019b), 4–21 und online: https://doi.org/10.11588/dco.2019.1.59356 (17.7.2021) und dies. (2017), 17–48, online: https://books.ub.uni-heidelberg.de/propylaeum/catalog/book/257?lang=en (17.7.2021).

[22] Diese Methode ist ausführlich und mit Literaturhinweisen beschrieben in: Schubert (2020), 305–327.

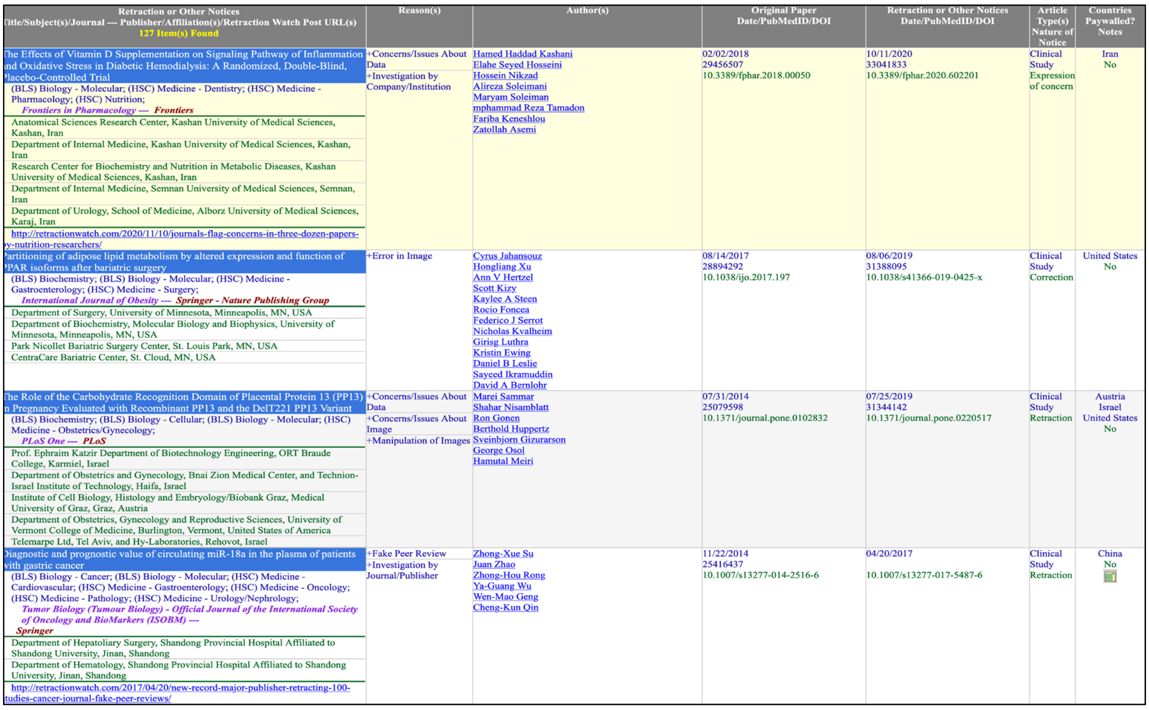
[23] Die Buchstaben N-Gramme (Buchstaben N=3) wurden mit StyloAH (M. Eders Stylo mit der Programmerweiterung durch H. Kahl: v.0.7.4.5: https://github.com/ecomp-shONgit/stylo [17.7.2021]) durchgeführt: Die Visualisierung erfolgte mit Gephi 0.9.1 (https://gephi.org/) im Layout OpenOrd. Zu OpenOrd: Martin et al. (2011), online: https://github.com/gephi/gephi/wiki/OpenOrd (17.7.2021).

[24] Zur Unechtheit der Schrift *De virtutibus et vitiis*: Flashar (^{2}2004), 207 und 274f. Zu den bei Diogenes Laertius III 80ff. überlieferten *Divisiones*: Flashar (^{2}2004), 96f.

tersuchung als Beispiel für einen Weg der Fehleranalyse im Hinblick auf die Entwicklung einer Fehlerkritik dienen, um den Umgang mit Metadaten zu systematisieren. Denn Metadaten sind heute nicht nur strukturgebend, sondern sie sind auch zu strukturvorgebenden Daten geworden. Sie prägen die Informationen und erzeugen so selbst eine Kontextualisierung, wie eben die Aufnahme einer modernen Rekonstruktionsedition in ihrer Weiterwirkung in modernen Bibliotheksdatenbanken zeigt. Demgegenüber steht die Forderung, daß sich auch die Praxis der Metadaten den Bedingungen zu fügen hat, die für jede wissenschaftlich-kritische Praxis gelten: Jedes Metadatenmodell muß theoretisch begründet und kritisch systematisiert werden. Auf einer solchen Grundlage genügt nicht nur eine Visualisierung von Textanalysen dem wissenschaftlichen Anspruch, sondern ist überhaupt erst eine wissenschaftliche Arbeit mit Metadaten möglich. Eine Analyse wie die hier kurz referierte, muß die Grundlage für eine wissenschaftlich seriöse Implementierung von Metadaten sein und diese wiederum ist die Aufgabe einer kritischen Wissenschaftspraxis, die jedoch heutzutage in den bibliographischen Metadaten – insbesondere, wenn sie automatisch erzeugt werden – nicht vorliegt und so ihrerseits zu Fehlern bei der digitalen Auswertung dieser Daten beiträgt.

II. Auf dem Weg zu einer Praxis konstruktiver Fehlerkultur

Seit einigen Jahren ist zu beobachten – empirisch belegt in einer zeitlichen Korrelation zum Fortschreiten der Digitalisierung – daß die Anzahl zurückgezogener Aufsätze und Publikationen ansteigt: In der Mehrzahl geschehen diese Zurücknahmen auf Grund experimenteller und unbeabsichtigter Fehler, es handelt sich also bei weitem nicht nur um Plagiate. Die Biochemie und Molekularbiologie, Zellbiologie und Onkologie führen das Feld der zurückgenommenen Artikel an. Dazu passt die Beobachtung, daß Journale mit sehr hohem Impact ebenfalls in den vordersten Reihen dieser Rücknahmen zu finden sind.[25] Die Datenbank Retraction Watch gibt einen sehr zeitnahen Überblick der zurückgezogenen Artikel, die nach Disziplinen aufgeschlüsselt sind.[26] Für den Bereich Molecular Biology verzeichnet die Datenbank mit Datum v. 17.7.2021 insg. 127 zurückgezogene Artikel:



Retraction or Other Notices Title/Subject(s)/Journal --- Publisher/Affiliation(s)/Retraction Watch Post URL(s) 127 Item(s) Found	Reason(s)	Author(s)	Original Paper Date/PubMedID/DOI	Retraction or Other Notices Date/PubMedID/DOI	Article Type(s) Nature of Notice	Countries Paywalled? Notes
The Effects of Vitamin D Supplementation on Signaling Pathway of Inflammation and Oxidative Stress in Diabetic Hemodialysis: A Randomized, Double-Blind, Placebo-Controlled Trial (BLS) Biology - Molecular; (HSC) Medicine - Dentistry; (HSC) Medicine - Pharmacology; (HSC) Nutrition; *Frontiers in Pharmacology --- Frontiers* Anatomical Sciences Research Center, Kashan University of Medical Sciences, Kashan, Iran Department of Internal Medicine, Kashan University of Medical Sciences, Kashan, Iran Research Center for Biochemistry and Nutrition in Metabolic Diseases, Kashan University of Medical Sciences, Kashan, Iran Department of Internal Medicine, Semnan University of Medical Sciences, Semnan, Iran Department of Urology, School of Medicine, Alborz University of Medical Sciences, Karaj, Iran http://retractionwatch.com/2020/11/10/journals-flag-concerns-in-three-dozen-papers-by-nutrition-researchers/	+Concerns/Issues About Data +Investigation by Company/Institution	Hamed Haddad Kashani Elahe Seyed Hosseini Hossein Nikzad Alireza Soleimani Maryam Soleiman mphammad Reza Tamadon Fariba Keneshlou Zatollah Asemi	02/02/2018 29456507 10.3389/fphar.2018.00050	10/11/2020 33041833 10.3389/fphar.2020.602201	Clinical Study Expression of concern	Iran No
Partitioning of adipose lipid metabolism by altered expression and function of PPAR isoforms after bariatric surgery (BLS) Biochemistry; (BLS) Biology - Molecular; (HSC) Medicine - Gastroenterology; (HSC) Medicine - Surgery; *International Journal of Obesity --- Springer - Nature Publishing Group* Department of Surgery, University of Minnesota, Minneapolis, MN, USA Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, USA Park Nicollet Bariatric Surgery Center, St. Louis Park, MN, USA CentraCare Bariatric Center, St. Cloud, MN, USA	+Error in Image	Cyrus Jahansouz Hongliang Xu Ann V Hertzel Scott Kizy Kaylee A Steen Rocio Foncea Federico J Serrot Nicholas Kvalheim Girisg Luthra Kristin Ewing Daniel B Leslie Sayeed Ikramuddin David A Bernlohr	08/14/2017 28894292 10.1038/ijo.2017.197	08/06/2019 31388095 10.1038/s41366-019-0425-x	Clinical Study Correction	United States No
The Role of the Carbohydrate Recognition Domain of Placental Protein 13 (PP13) in Pregnancy Evaluated with Recombinant PP13 and the DelT221 PP13 Variant (BLS) Biochemistry; (BLS) Biology - Cellular; (BLS) Biology - Molecular; (HSC) Medicine - Obstetrics/Gynecology; *PLoS One --- PLoS* Prof. Ephraim Katzir Department of Biotechnology Engineering, ORT Braude College, Karmiel, Israel Department of Obstetrics and Gynecology, Bnai Zion Medical Center, and Technion-Israel Institute of Technology, Haifa, Israel Institute of Cell Biology, Histology and Embryology/Biobank Graz, Medical University of Graz, Graz, Austria Department of Obstetrics, Gynecology and Reproductive Sciences, University of Vermont College of Medicine, Burlington, Vermont, United States of America Telemarpe Ltd, Tel Aviv, and Hy-Laboratories, Rehovot, Israel	+Concerns/Issues About Data +Concerns/Issues About Image +Manipulation of Images	Marei Sammar Shahar Nisamblatt Ron Gonen Berthold Huppertz Sveinbjorn Gizurarson George Osol Hamutal Meiri	07/31/2014 25079598 10.1371/journal.pone.0102832	07/25/2019 31344142 10.1371/journal.pone.0220517	Clinical Study Retraction	Austria Israel United States No
Diagnostic and prognostic value of circulating miR-18a in the plasma of patients with gastric cancer (BLS) Biology - Cancer; (BLS) Biology - Molecular; (HSC) Medicine - Cardiovascular; (HSC) Medicine - Gastroenterology; (HSC) Medicine - Oncology; (HSC) Medicine - Pathology; (HSC) Medicine - Urology/Nephrology; *Tumor Biology (Tumour Biology) - Official Journal of the International Society of Oncology and BioMarkers (ISOBM) --- Springer* Department of Hepatoliary Surgery, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, Shandong Department of Hematology, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, Shandong http://retractionwatch.com/2017/04/20/new-record-major-publisher-retracting-100-studies-cancer-journal-fake-peer-reviews/	+Fake Peer Review +Investigation by Journal/Publisher	Zhong-Xue Su Juan Zhao Zhong-Hou Rong Ya-Guang Wu Wen-Mao Geng Cheng-Kun Qin	11/22/2014 25416437 10.1007/s13277-014-2516-6	04/20/2017 10.1007/s13277-017-5487-6	Clinical Study Retraction	China No ▤

Abb. 4: http://retractiondatabase.org mit den Suchparametern Biology – Molecular/Clinical Study (17.7.2021).

[25] Dollfuß (2015): https://www.egms.de/static/de/journals/mbi/2015-15/mbi000336.shtml#ref2 (17.7.2021).

[26] http://retractiondatabase.org (17.7.2021).

Es geht im Kern darum, daß Fehler gemacht werden, und vor allem, daß Ergebnisse nicht reproduzierbar sind.[27] Daher stellt sich die Frage, wie man mit diesem Befund umgehen soll. Für die Digital Humanities hat Nan Z. Da in ihrem Beitrag gerade diese Fehlerhaftigkeit und mangelnde Reproduzierbarkeit kritisiert. Sie geht jedoch noch weiter: Ihrer Ansicht nach werden Fehlklassifizierungen sogar zum Objekt des Interesses gemacht, Ungenauigkeiten und Sonderfälle würden theoretisiert, und alles dies werde Grundlage für Forschungsförderung und Publikationen.[28]

Eine sehr nachdrückliche Gegenposition von Seiten der Digital Humanities ist dazu von Christof Schöch formuliert worden,[29] der den Anspruch an die Replizierbarkeit folgendermaßen ansetzt: „The typology describes the relationship between an earlier study and its replication in terms of four key variables: the research question, the method of analysis (including the implementation of that method) and the dataset used."[30] Etwas anders setzt Fotis Jannidis an,[31] der, ausgehend davon, daß die Literary Studies schon immer eine empirisch-quantitative Seite gehabt hätten, in diesem Feld lediglich einen inhärenten Wechsel zweier Methoden, der hermeneutischen und der quantitativen, konstatiert, die sich im Grunde genommen auch überlappen würden.[32] Diese hier sehr kurz und auch nur schematisch skizzierten Überlegungen zeigen, daß das Problem in den Digital Humanities erkannt und auch angegangen wird, wenngleich die Diskussion doch noch sehr dem Abwehrmodus verhaftet zu sein scheint.

Ganz anders ist die Position der einflußreichsten Forschungsförderorganisation in Deutschland im Hinblick auf die Replizierbarkeitsproblematik.[33] In ihrer diesbezüglichen Stellungnahme aus dem Jahr 2017 schließt die Deutsche Forschungsgemeinschaft (DFG) ausdrücklich alle Wissenschaftsbereiche ein und hält fest: „Replikation ist ein sehr wichtiges Verfahren zur Prüfung experimentalwissenschaftlich und quantitativ begründeter empirischer Wissensansprüche in der Medizin, in den Natur-, Lebens-, Ingenieur- sowie den Sozial- und Verhaltenswissenschaften und auch den Geisteswissenschaften." Interessant ist, daß die DFG hier zwar die Geisteswissenschaften einbezieht, jedoch mit der einschränkenden Begrenzung auf die quantitativ begründeten, empirischen Wissenschaften. Das 2018 veröffentlichte

[27] Eine einleuchtende Begriffsbestimmung zu Replikation und Reproduktion findet sich in DFG (2018), 12: (https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2018/180507_stellungnahme_replizierbarkeit_sgkf.pdf, abgerufen 17.7.2021):
„Die Begriffsbildung der Wörter Replikation, Replizierbarkeit bzw. der ebenfalls oft verwendeten Begriffe Reproduktion, Reproduzierbarkeit sowie die Abgrenzungen der Bedeutung zwischen diesen Ausdrücken ist noch nicht abgeschlossen. Die Begriffe werden auch im angelsächsischen Sprachraum nicht einheitlich verwendet.
In diesem Papier werden nur die Begriffe Replikation und Replizierbarkeit verwendet, mit folgender Bedeutung:
Replikation ist die Wiederholung einer Untersuchung/eines Experiments/einer Studie, die den Anspruch auf Wiederholbarkeit erhebt, meint aber auch allgemein die ergebnisoffene Möglichkeit, etwas zu wiederholen bzw. ein Experiment noch einmal durchführen zu können.
Replizierbarkeit meint entsprechend die Möglichkeit bzw. Fähigkeit, Ergebnisse innerhalb des Fehlerrahmens wiederholend zu bestätigen."

[28] Da (2019a): „CLS [sc. Computational Literary Studies, C.S.] routinely relies on these concepts to provide plausible explanations or theoretical motivations for results that are nothing more than a description of the data. In their project on The Sorrows of Young Werther, for example, Andrew Piper and Mark Algee-Hewitt compared a standard visualization of the repetition of 91 words in Goethe's oeuvre with theoretical paradigms as different as those of Gilles Deleuze, Alain Badiou, Bruno Latour, and Michel Foucault."

[29] Schöch et al. (2020), online: https://hcommons.org/deposits/item/hc:30439/ (17.7.2021).

[30] Schöch, Contribution 1 in Schöch et al. (2020) (wie Anm. 29), S. 2 der PDF-Version.

[31] Jannidis, Contribution 3 in Schöch et al. (2020) (wie Anm. 29), S. 6 der PDF-Version.

[32] Da die klassische Philologie sich bereits seit dem 19. Jahrhundert solcher empirisch-quantitativen Analysen bedient hat und auch die Alte Geschichte, bspw. im Bereich der Epigraphik und Numismatik, mit ähnlichen Ansätzen arbeitet, ist Jannidis in dem Punkt natürlich zustimmen, daß es solche empirischen Studien auch in geisteswissenschaftlichen Fächern schon lange gibt.

[33] DFG (2017) zur Replizierbarkeit von Forschungsergebnissen, online: https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf (17.7.2021).

DFG-Papier „Replizierbarkeit von Ergebnissen in der Medizin und Biomedizin. Stellungnahme der Arbeitsgruppe ‚Qualität in der Klinischen Forschung' der DFG-Senatskommission für Grundsatzfragen in der Klinischen Forschung" (2018) bezieht ausdrücklich nur die Medizin und Biomedizin ein.[34]

Aufschlußreich ist es, dazu einige Zitate aus der Veröffentlichung der DFG zur Frage der Replizierbarkeit von Forschungsergebnissen anzusehen.[35] Die drei Kernaussagen der Stellungnahme aus 2017 sind:

- „Replizierbarkeit ist kein generelles Kriterium wissenschaftlicher Erkenntnis."
- „Nicht-Replizierbarkeit ist kein genereller Falsifikationsbeweis."
- „Die Feststellung der Replizierbarkeit oder Nicht-Replizierbarkeit eines wissenschaftlichen Ergebnisses ist ihrerseits ein wissenschaftliches Ergebnis."

Die Frage der Replizierbarkeit ist, vor dem Hintergrund der massiven Kritik an den Digital Humanities, von grundsätzlicher Bedeutung für ihre Position im Verhältnis zu den klassischen Geisteswissenschaften. Es ist nicht zu bestreiten, daß ein hermeneutisches Vorgehen sich nicht replizieren läßt in dem Sinn, daß unter gleichen Bedingungen (wie etwa im Labor, bei Experimenten, bei empirischen und/oder quantitativen Studien) ein Ergebnis sozusagen wiederholbar ist. Vielmehr müssen das Ergebnis und Verfahren des hermeneutischen Vorgehens in ihrer Sinnhaftigkeit überzeugen und erhalten dadurch den gleichen wissenschaftlichen Stellenwert wie die Wiederholbarkeit durch Replikation.

Im Unterschied dazu ist bei Ergebnissen, die mit den Methoden der Digital Humanities erzielt worden sind, Replikation bzw. Reproduzierbarkeit möglich, denn damit ist die Wiederholung einer Analyse mit dem gleichen Datensatz und den gleichen Methoden gemeint. Hierfür könnten die von der DFG 2018 für den Bereich der Medizin und Biomedizin formulierten Grundsätze, die als Voraussetzungen der Replizierbarkeit dienen, durchaus übertragen werden:[36]

- „Validität von Modellen und Standardisierung von Methoden"
- „Adäquate statistische Planung"
- „Sorgfältiges Management von Forschungsdaten und Materialien"
- „Umfassende Darstellung der Methoden und Analysen"

Insofern sollte Replizierbarkeit in den Digital Humanities als hinreichend betrachtet werden, wenn Verfahren und Ergebnis auf diesen Voraussetzungen beruhen. Wenn diese Form der Replizierbarkeit nicht gegeben und auch noch nicht als Standard etabliert ist, dürften auch weiterhin Zweifel an der Validität aufkommen, und – wie die Diskussion, die hier skizziert ist, deutlich zeigt, – die Ergebnisse als fehlerhaft klassifiziert werden. Hinzu kommt darüber hinaus die kritische Begründung der konzeptionellen Voraussetzungen, die aber von der eigentlichen Replizierbarkeit unterschieden werden sollte und den hier genannten Voraussetzungen der Replizierbarkeit vorausgehen muss.[37]

Damit stellt sich die grundlegende Frage, wie angesichts der laufenden Auseinandersetzungen mit Fehlerhaftigkeit in den Digital Humanities umgegangen werden soll. Wenn man bspw. in der Alten Geschichte mit dem klassischen Methodeninstrumentarium arbeitet, dann verwendet man bei der Arbeit mit Textquellen Editionen mit einem kritischen Apparat, der auf der Sichtung der Handschriften durch

[34] DFG (2018): https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2018/180507_stellungnahme_replizierbarkeit_sgkf.pdf (17.7.2021).

[35] DFG (2017) zur Replizierbarkeit von Forschungsergebnissen, online: https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf (17.7.2021).

[36] DFG (2018) zu Replizierbarkeit von Ergebnissen in der Medizin und Biomedizin: https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2018/180507_stellungnahme_replizierbarkeit_sgkf.pdf, S. 8 (17.7.2021).

[37] Anders Jannidis, Contribution 3 in Schöch et al. (2020).

die Editoren beruht. Man stellt Quellenreferenzen und verwendete Literatur zusammen und versucht dies in der Regel in vollständiger Weise, um die Argumente zu belegen und so auch zu plausibilisieren. Unterlaufen Fehler oder vielleicht sogar nur vermeintliche Fehler, so ist dies Ausgang eines wissenschaftlichen Diskurses, der im Fach geführt wird. Dabei kommt es durchaus zu Schärfen gegenüber denjenigen, denen man Fehler nachweist oder auch nachzuweisen glaubt, jedoch wird niemand das ganze Methodeninstrumentarium oder das Fach selbst infrage stellen.

In den Digital Humanities beobachten wir eine andere Art der Diskussion um Fehler, wie die zitierten Beiträge, aber auch viele andere zeigen: Hier wird dem ganzen Bereich eine grundsätzliche Schwäche und Fehlerhaftigkeit attestiert, denn Fehler, Ungenauigkeiten und Nicht-Wiederholbarkeit werden als ‚Schuld‘ eines falschen Vorgehens klassifiziert. Dies ist ganz offensichtlich der Versuch, die wissenschaftliche Entwicklung der Digital Humanities in ihrem Methodenbereich mit der Figur einer Schuldkultur zu verbinden.

Demgegenüber sollte man für die zweifelsfrei sichtbaren Probleme und Fehler eine konstruktive Fehlerkultur für die Digital Humanities entwickeln und offensiv mit den Fehlern umgehen. Daher wird hier ein methodisches Vorgehen vorgeschlagen, durch das eine konstruktive Fehlerkultur in den Digital Humanities ermöglicht und eine Brücke zu den klassisch hermeneutischen Arbeitsweisen gebaut würde.

Grundsätzlich ist zuallererst eine Taxonomie der Fehler nötig,[38] denn je nach Arbeitsphase oder -stand sind die Fehler andere und ist auch anders damit umzugehen. Die Fehler müssen klassifiziert werden nach Grundannahmen, Experimentphasen und Komplexitätskontext.

Darauf muß eine Fehlerkritik aufsetzen, die die Fehler entsprechend ihrer Klassifikation einordnet und in Iterationen so integriert, so daß die Veränderungen in Daten, Metadaten und Parameter immer – auch nachträglich – zugeordnet und transparent gemacht werden können.

Schließlich muß das Ergebnis selbst in jeder Hinsicht transparent gemacht werden, so daß der Weg dahin jederzeit nachverfolgt werden kann. Am Ende muß auch Ergebnisstabilität erzielt werden, so daß die Digital Humanities aus dem Projektstadium zum „Dauerbetrieb“ kommen.

Letztlich würde dies auf die Notwendigkeit der Standardisierung der Methoden führen und dies könnte mit der Einführung von SOPs (Standard Operating Procedures) in den Digital Humanities erreicht werden, die sie vergleichbar mit anderen technischen Bereichen machen. Somit müssen nicht nur die einzelnen Schritte des Preprocessing dokumentiert werden, sondern vor allem müssen die Voraussetzungen und die der Datenaufbereitung zugrundeliegende Systematik offengelegt und in ein Verhältnis zur jeweiligen, fachspezifischen Wissensordnung gesetzt werden.

Dies sind anspruchsvolle Anforderungen, die aber den Digital Humanities die Augenhöhe sowohl mit anderen technisch orientierten Fächern wie vor allem mit den klassischen Geisteswissenschaften ermöglichen werden.

[38] Schöch in Schöch, Contribution 1 in Schöch et al. (2020), S. 2f. (der PDF-Version) schlägt eine Typologie der Replikation vor, für die er auch eine Systematik visualisiert. M.E. ist es jedoch notwendig, bevor die Digital Humanities dies angehen können, die Fehler kritisch zu analysieren und dann erst die Replikationssystematik zu entwickeln.

Literatur

Da (2019a): Nan Z. Da, The Digital Humanities Debacle, The Chronicle of Higher Education 2019. Online: https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986

Da (2019b): Nan Z. Da., The Computational Case against. Computational Literary Studies, Critical Inquiry 45/3 (Spring 2019), 601–639. Online: https://www.journals.uchicago.edu/doi/abs/10.1086/702594?journalCode=ci

DFG (2017): DFG-Positionspapier zur Replizierbarkeit von Forschungsergebnissen, Stellungnahme der Deutschen Forschungsgemeinschaft, Bonn 2017. Online: https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf

DFG (2018): DFG Positionspapier: „Replizierbarkeit von Ergebnissen in der Medizin und Biomedizin. Stellungnahme der Arbeitsgruppe ‚Qualität in der Klinischen Forschung' der DFG-Senatskommission für Grundsatzfragen in der Klinischen Forschung", Bonn 2018. Online: https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2018/180507_stellungnahme_replizierbarkeit_sgkf.pdf

Diemke (2020): J. Diemke, Alkibiades, Pyrrhos und Alexander: Eine Untersuchung zu Emotionen und Gewalt in den Viten Plutarchs unter Verwendung digitaler Methoden, DCO 6,2 (2020), 57–74. Online: https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/77663/71565

Dollfuß (2014): H. Dollfuß, Analyse zurückgezogener Publikationen in der bibliografischen Datenbank Web of Science von 2004 bis 2014, MS Med Bibl Inf 2015;15(1–2):Doc09. Online: https://www.egms.de/static/de/journals/mbi/2015-15/mbi000336.shtml#ref2

Flashar (2004): H. Flashar, das Werk des Aristoteles, in: H. Flashar (Hrsg.), Die Philosophie der Antike Bd. 3: Ältere Akademie. Aristoteles. Peripatos, Basel ²2004.

Juola (2015): P. Juola, The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions, Digital Scholarship 30 (2015), Supplement 1, 101–113.

Kirby (2019): J. S. Kirby, J. S. (2019), How NOT to create a digital media scholarship platform: the history of the Sophie 2.0 project, IASSIST Quarterly, 42,4 (2019). Online: https://doi.org/10.29173/iq926

Krämer (2018): S. Krämer, Der ‚Stachel des Digitalen' – ein Anreiz zur Selbstreflexion in den Geisteswissenschaften? Ein philosophischer Kommentar zu den Digital Humanities in neun Thesen, DCO 4,1 (2018), 5–11. Online: https://doi.org/10.11588/dco.2017.0.48490

Martin et al. (2011): W. Martin, M. Brown / R. Klavans / K. Boyack, OpenOrd: An Open-Source Toolbox for Large Graph Layout, 2011, Proceedings of SPIE – The International Society for Optical Engineering SPIE 2011. Online: https://github.com/gephi/gephi/wiki/OpenOrd

Rautenberg (2019): J. Rautenberg, Negation in Platons Sophistes und die Grenzen automatisierter Paraphrasensuche, in: Schubert, Ch. et al. (Hrsgg.): Platon Digital: Tradition und Rezeption, Heidelberg 2019, 111–123 (Digital Classics Books, Band 3). Online: https://doi.org/10.11588/propylaeum.451

Riess (2019): W. Riess, A Digital Analysis of Maritime Acts of Violence Committed by Alcibiades as Described by Thucydides, Xenophon, and Plutarch, DCO 5,2 (2019), 4–27. Online: https://doi.org/10.11588/dco.2019.2.72018

Riess (2020): W. Riess, Prolegomena zu einer digitalen althistorischen Gewaltforschung: Gewaltmuster bei Solon, Alkibiades und Arat im Vergleich, Klio 102 (2020), 445–473.

Rizvi (2018): P. Rizvi, The interpretation of Zeta test results, Digital Scholarship in the Humanities 34/2 (2018), 401–418. Online: http://doi.org/10.1093/llc/fqy038

Schöch et al. (2020): Chr. Schöch et al., Replication and Computational Literary Studies, Panel at the Digital Humanities Conference 2020 (DH2020), Ottawa, Canada, July 20–25, 2020. Online: https://hcommons.org/deposits/item/hc:30439/ und https://dh2020.adho.org/

Schäfer (2019): Ch. Schäfer, Die Kontrolle des Meeres: Alkibiades und die Sizilische Expedition, DCO 5,1 (2019), 22–33 und online: https://doi.org/10.11588/dco.2019.1.60564

Schubert / Weiß (2015): Ch. Schubert / A. Weiß, Die Hypomnemata bei Plutarch und Clemens: Ein Textmining-gestützter Vergleich der Arbeitsweise zweier ‚Sophisten', Hermes 143 (2015), 447–471.

Schubert (2017): Ch. Schubert, Die Arbeitsweise Iamblichs im Protreptikos, in: S. Brandt / Ch. Schubert: Der Protreptikos des Iamblich: Rekonstruktion, Refragmentisierung und Kontextualisierung mit Textmining, Digital Classics Books, Heidelberg 2017, 17–48. Online: https://books.ub.uni-heidelberg.de/propylaeum/catalog/book/257

Schubert (2018): Ch. Schubert, Eine Thukydides-Paraphrase in der Totenrede des Tiberius auf Augustus: Cassius Dios Sichtweise des augusteischen Prinzipats, Antike und Abendland 64 (2018), 79–92.

Schubert (2019a): Ch. Schubert, Plädoyer für eine Fehlerkultur in den Digital Humanities, DCO 5,2 (2019). Online: https://doi.org/10.11588/dco.2019.2.72004

Schubert (2019b): Ch. Schubert, Visualisierung von Textdaten: Die Falle der Metadaten am Beispiel von Iamblichs Protreptikos, DCO 5,1 (2019), 4–21. Online: https://doi.org/10.11588/dco.2019.1.59356

Schubert (2020): Ch. Schubert, Zur Standortbestimmung des Digitalen in den Altertumswissenschaften. Textanalyse am Beispiel des Corpus Hippocraticum und des hippokratischen Eides, Gymnasium 127 (2020), 305–327.

Thomas (2016): William G. Thomas III, The Promise of the Digital Humanities and the Contested Nature of Digital Scholarship in: Susan Schreibman, Ray Siemens, John Unsworth (Hrsgg.), A New Companion to Digital Humanities, Malden / Oxford ^{2}2016, 524–537.

Van Zundert (2016): J. J. van Zundert, Screwmeneutics and Hermenumericals: the Computationality of Hermeneutics, in: S. Schreibman / R. Siemens / J. Unsworth (Hrsgg.), A New Companion to Digital Humanities, Malden/ Oxford 2016, 331–347. Online: http://onlinelibrary.wiley.com/doi/10.1002/9781118680605.ch23/summary

Warnking (2015): P. Warnking, Der römische Seehandel in seiner Blütezeit: Rahmenbedingungen, Seewege, Wirtschaftlichkeit, Rahden / Westf. 2015.

Warnking (2016): P. Warnking, Roman Trade Routes in the Mediterranean Sea: Modelling the routes and duration of ancient travel with modern offshore regatta software, in: Ch. Schäfer (Hrsg.), Connecting the Ancient World. Mediterranean Shipping, Maritime Networks and their Impact, Rahden / Westf. 2016, 45–90.

Internetseiten

DNB:
https://www.dnb.de/DE/Professionell/Standardisierung/Standards/standards_node.html
https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
http://d-nb.info/gnd/4305003-7

Gephi 0.9.1:
https://gephi.org/

Infowiss:
https://wiki.infowiss.net/Stoppwort

Retraction Database:
http://retractiondatabase.org

StyloAH v.0.7.4.5:
https://github.com/ecomp-shONgit/stylo

TLG-Datenbank:
http://stephanus.tlg.uci.edu/

Webportal Digital Plato:
https://digital-plato.org/

Webportal eAQUA:
www.eaqua.net/

Zusammenstellung der Antworten auf Nan Z.Da:
https://demo.hedgedoc.org/s/rJ_YoK_cH

Abbildungsnachweise

Abb. 1: Stoppwortliste, Ausschnitt aus: https://wiki.infowiss.net/Stoppwort
Abb. 2: Katalog der DNB, Ausschnitt aus: http://d-nb.info/gnd/4305003-7
Abb. 3: Ch. Schubert
Abb. 4: Retraction Database, Ausschnitt aus: http://retractiondatabase.org

Autoreninformation[39]

Prof. em. Dr. Charlotte Schubert

Universität Leipzig
Beethovenstr. 15
04107 Leipzig
Tel: 0341/9737071
E-Mail: schubert@uni-leipzig.de

[39] Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren. Alle Inhalte dieses Beitrages unterstehen, soweit nicht anders gekennzeichnet, der Lizenz CC BY 4.0.

##DIGITAL CLASSICS ONLINE##

Pain and the Body in *Corpus Hippocraticum*: A Distributional Semantic Analysis

Vojtěch Linka, Vojtěch Kaše

Abstract: The authors of the medical treatises collected in *Corpus Hippocraticum* often mention pain, its qualities and origin. At the same time, however, they do not provide any explicit definition or theory of pain, of its nature and of relation to other important aspects of Hippocratic medicine. Moreover, they employ at least four word-families which are commonly suggested to denote pain in ancient Greek. This encourages modern researchers to ask how do these four pain-words semantically differ and to what extent are they based on a shared notion of pain. In this article, we attempt to answer these questions by analysing the corpus employing several computational text analysis methods, especially by employing a distributional semantic modelling approach. Our results reveal a close association between some of these pain-words, bodily parts and pathological states. The results are further compared with findings obtained through the traditional close reading of the sources.

Introduction[1]

Alleviating pain and taking away the cause of suffering is one of the maxims of Hippocratic medicine.[2] The authors of Hippocratic writings often mention pain; we read about its quality, location, causes and relation to illness. What, though, is pain? There is no explicit definition of pain in *Corpus Hippocraticum* (= *CH*); there is no treatise on its nature. While some scholars assume there might be a unified conception of pain in *CH*,[3] other researchers boldly claim there is no such thing.[4] In this article, we evaluate these propositions by combining insights obtained by the application of some computational text analyses methods on the corpus as a whole (i.e. *distant reading*), with insights based on the detailed interpretation of selected passages from the corpus (i.e. *close reading*).[5]

In particular, we focus on the semantics of a selection of words which are commonly suggested to denote pain in *CH*, namely four word-families: πόνο*, ὀδύν*, ἄλγ*, λύπ*. We are especially interested in the typical contexts in which these pain-words occur, how these words are related to each other and what other words are most closely semantically associated with them.

We begin with the Materials and Methods section, in which we offer an overview of the data we use for the computational text analysis parts of the article. We describe the procedures which were conducted to make the texts suitable for these analyses and subsequently introduce all methodological steps, especi-

[1] This work was financially supported by Charles University Grant Agency, project no. 78120, entitled "Aristotle and Hippocrates on Pain", implemented at the Faculty of Arts of Charles University.

[2] Hippocr. *Med. Vet*. 3.35–40; *Vict*. 15.5–6; *De arte* 3.5.

[3] King (1999), 269–286.

[4] Horden (1999), 295–315.

[5] The phrases distant reading and close reading were coined by Franco Moretti and have since come to be used widely in digital humanities literature. See Moretti (2013); Underwood (2017), 1–12; Jänicke et al. (2015), 1–21.

ally concerning the Distributional Semantic analysis. We continue with the Results section which builds on these methods. Finally, in the Discussion, we evaluate the results obtained by the computational text analysis methods against observations based on the close reading of individual texts and passages from the corpus.

Materials and Methods

Textual corpus and its preprocessing

The computational text analyses included in this article are based on a corpus of Hippocratic texts retrieved from the Lemmatised Ancient Greek Texts dataset (LAGT).[6] LAGT combines two open-source corpora of ancient Greek texts: the Canonical Greek Literature dataset from the Perseus Digital Library[7] and the First Thousand Years of Greek dataset of the Open Greek & Latin project.[8] Both datasets are publicly available on Github and Zenodo under Creative Commons Attribution licences, which makes them suitable for further reuse within any large-scale computational text analysis project[9] or within LAGT and, subsequently, in our article.[10] Further, the works in LAGT employ a canonical reference system based on the CITE architecture,[11] which makes identification of any work in the dataset very straightforward.[12]

Within LAGT, the textual data from the Perseus Digital Library and the First Thousand Years of Greek project are subjected to several standard text preprocessing procedures, namely tokenization, POS-tagging and lemmatisation. To provide a better understanding of the subsequent analyses, it will be useful to briefly describe how they are implemented within LAGT. Tokenization is the procedure of splitting textual data into their constitutive elements, called tokens. Thus, within LAGT, each work is first divided into sentences and subsequently each sentence into words. In the next step, each token (i.e. word) is coupled with a POS tag[13] and a lemma.[14] Assignment of a POS tag works probabilistically as an output of a neural network model which has been previously trained on manually annotated ancient Greek sentences.[15] Subsequently, the lemmatisation works deterministically, trying to find a suitable word-

[6] Kaše (2021).

[7] Cerrato et al. (2020).

[8] Crane et al. (2020).

[9] E.g. Koentges (2020).

[10] Because of relying on LAGT, some of our calculations might be slightly different from the ones we could obtain by employing other digital editions, namely from Thesaurus Linguae Graecae (TLG). For instance, the LAGT dataset contains a substantially shorter version of the Hippocratic treatise *Epidemiae*. This is due to the fact that the Perseus version of this work relies on Loeb's edition from 1923, which treats only books I and III as substantial representatives of epidemic medicine and does not contain books II and IV–VII.

[11] Blackwell / Smith (2019).

[12] Thus, the Hippocratic texts might be easily extracted using the CTS URN for author "tlg0627".

[13] POS stands for part of speech. In the case of LAGT, the POS-tagging has the form of a coarse-grained analysis, which means that it assigns to a word only the part of speech itself (e.g. noun, verb, adjective, conjunction etc.) and not other morphological features such as gender, number, or tense (i.e. fine-grained analysis). For the POS tag categories, see https://universaldependencies.org/u/pos/ (Last access 10.07.2021).

[14] Lemma is the dictionary form of a word. Thus, in the case of a verb, it is 1st sing. pres. ind. act. (e.g. δοκέω).

[15] I.e., the model tries to predict the POS tag of a word by drawing on the structure of the current sentence and comparing it with sentences that the model encountered during its training. See the LAGT repository for more details.

form-lemma pair within the Greek part of the Morpheus Dictionary.[16] Having the words coupled with their POS tags makes it possible to filter texts according to them and focus only on lemmatised words coupled with certain POS tag categories. Since LAGT is primarily designed for semantic analysis, it returns lemmatised versions of the texts containing only words tagged as nouns, proper names, adjectives and verbs.[17] Being aware that the POS tagging and lemmatisation are semi-automatic processes, we should not be surprised that they are also prone to errors. Therefore, our data may still contain a negligible amount of incorrectly POS-tagged, improperly lemmatised or completely un-lemmatised words.[18] Despite this fact, it seems that this limitation does not bias the overall results of our analyses.

All computational text analyses introduced below have been implemented using the Python 3 programming language.[19] Since we aim to make our analyses fully reproducible and our code reusable by other scholars, all the data and the whole code used in this article are accessible via a Zenodo repository,[20] to which we occasionally refer below for details and supplementary data and figures.

Document distances

To obtain a general overview of the corpus and the relationship between individual documents, we firstly generated a document-term matrix, with rows representing individual works in the corpus and columns representing a subselection of words used in the corpus. The cell values within the matrix represent frequencies of these words across the works within the corpus. In particular, the selected words are words appearing in at least 10% of works in the corpus. This forms a set of 2,033 unique words. The rows of this matrix have been subsequently treated as vectors, expressing positions of points in a multidimensional space, with the number of dimensions equal to the number of words. Thus, we obtained a set of 52 vectors within a space with 2,033 dimensions. Having the data in this form, we can calculate distances between the vectors by measuring and inverting their cosine similarity. This way we obtain a matrix expressing distance between any two works within the corpus, with works sharing a larger proportion of words being less remote to each other than works employing less overlapping vocabulary. This distance matrix could be finally projected into a 2-dimensional space by using t-distributed Stochastic Neighbor Embedding (tSNE)[21] and plotted as a scatter plot (see Fig. 1 below).[22]

Pain-words in context

As we have already mentioned, when it comes to the concept of pain in *CH*, we have to deal with at least four word-families: πόνο*, ὀδύν*, ἄλγ*, λύπ*. Each of these word families combines several words, which we usually have in the lemmatised form. Thus, for instance, the most common lemmata from the ἄλγ* family appear to be the noun ἄλγημα and the verb ἀλγέω, with 141 and 84 instances respectively. However, in our corpus, there is also a significant number of instances of un-lemmatised words which

[16] Crane (1991).

[17] For a similar approach see Svärd et al. (2020), 470–502. This approach differs from computational stylometry, which commonly focuses on the usage of conjunctions, prepositions etc., which usually capture any difference in a style very distinctly. Cf. Koentges (2020), 211–41.

[18] The accuracy (proportion of correctly annotated words from a text) of the POS-tagger and the lemmatiser is between 87 and 97 %, depending on the genre of the text.

[19] Rossum / Drake (2009).

[20] Kaše / Linka (2021).

[21] van der Maaten / Hinton (2008), 2579–2605.

[22] For details, see Kaše / Linka (2021), scripts/3_OVERVIEW+WORK-DISTANCES.ipynb (Last access 30.08.2021).

are not covered by the database we used for lemmatisation, like ἀλγεῦντα. We used regular expressions[23] to capture all these word forms and replaced them with a unified word pattern consisting of the word root and an asterisk: πόνο*, ὀδύν*, ἄλγ*, λύπ*.[24] In what follows, whenever we point out Greek pain-words, we refer to these word patterns.

Having the pain-words captured, we can focus on the context in which they appear. While the analysis of work distances treats the corpus as a list of individual works and each work as a list of words coupled with their frequencies, this analysis approaches the corpus as a list of sentences. For each pain-word, we firstly extract all sentences containing it. Subsequently, we compute term frequency (TF) for all words within these sentences. This measure gives us a general overview of the terms most commonly co-occurring with each of the pain-words. However, this measure does not distinguish between frequently appearing words in the sentences containing the pain-words, as they are semantically associated with them, and words frequently appearing here because of their distribution over the corpus as a whole. To overcome this limitation, we weight the TF measurement using a TFIDF algorithm. TFIDF stands for term frequency-inverse document frequency. Inverse document frequency (IDF) is obtained by dividing the total number of documents by the number of documents containing the term. Typically, the IDF value is logarithmically normalized. TFIDF is then a multiplication of the two measures:

$$TFIDF \ = \ TF \times log_2(IDF)$$

In effect, the weighting by IDF proportionally reduces the TF values of context general words while increasing the values of context-specific words. Using this measure, for each of the pain-words, we are able to identify the words which are most typical for their context.

Distributional Semantics and PPMI[3]

A step further is to adopt some methods from the field of distributional (or vector) semantics. The term distributional semantics designates a broad palette of methods from the areas of natural language processing and computational linguistics inspired by the distributional hypothesis of meaning,[25] henceforth it is also called Distributional Semantic Modeling (DSM).[26] Since these methods usually transform words into vectors, some scholars use the designation vector semantics.[27] According to the distributional hypothesis, words that occur in similar contexts tend to have similar meanings. Thus, to capture the meaning of a word requires an analysis of words most frequently surrounding it. But, as we will see, this is only a starting point. DSM goes further and constructs matrices and vector representations for whole corpora, which are subsequently transformed and analysed using complex algorithms from linear algebra and statistics. However, to work properly, most of the DSM algorithms require very large textual data (typically at least 1 million words) to be trained on. In this respect, our corpus consisting of 171,332 words is rather small and therefore allows us to employ only certain distributional semantic models.[28]

[23] López / Romero (2014).

[24] We use the asterisk character (*) to mark that the word is a product of a regular expression match.

[25] For distribution hypothesis, see Harris (1954), 146–62.

[26] Lenci (2018), 151–71.

[27] For a basic overview of the most common algorithms, see Jurafsky / Martin (2020), 270–85.

[28] For instance, it has been demonstrated that the well-known word2vec model outperforms other methods when trained on 1 billion words of data. However, when trained on a smaller dataset, consisting of 1, 10 or 100 million words, it is outcompeted by much simpler models. For word2vec, see Mikolov et al. (2013), 3111–3119. For comparison of word2vec with other models, see Sahlgren / Lenci (2016); Altszyler et al. (2016).

In what follows, we employ a DSM approach combining Pointwise Mutual Information (PMI) and Singular Value Decomposition (SVD). In its basic version, PMI has the following form:

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)},$$ [28]

where x and y represent two words, *P(x,y)* their probability of appearing together within a predefined context within a corpus, and *P(x)* and *P(y)* their probabilities of appearing independently, i.e. their respective term frequencies within the corpus.[30] The ratio is subsequently normalized by a logarithm with base 2. However, a well-known problem associated with this measure is that it gives very high scores to word pairs involving infrequent words, as the denominator is rather small in such cases. Therefore, several of modifications of PMI have been proposed to overcome this limitation.[31] Here we employ the so-called PMI^3, which modifies the measure by cubing the *P(x,y)* value and which has been documented to produce reasonable results:

$$PMI^3(x,y) = log_2 \frac{P(x,y)^3}{P(x)P(y)}$$ [31]

Finally, since the fraction $\frac{P(x,y)^3}{P(x)P(y)}$ usually returns values lower than 1 and since *log*$_2$ for numbers smaller than 1 is a negative number (which might be confusing for a visual inspection), we finally create a PPMI^3 measure transposing the PMI^3 values to a scale from 0 to 1, with PPMI^3=0 for all word pairs with joint probability P(x,y) equal to 0 (i.e. for words which do not appear together at all) and the rest on a scale from 0.5 to 1, with PPMI^3=0.5 assigned to a word pair with the minimal PMI^3 value in total (but different from 0) and PPMI^3=1 assigned to a word pair with the maximal PMI^3 value in total.

Drawing on this, we can generate a PPMI^3 matrix by calculating the PPMI^3 value for each possible word pair of all words appearing in at least 5 works within the corpus. Such a matrix gives us straightforward access to weighted *first-order co-occurrence* (also called *syntagmatic association*) between any two words forming the matrix. Thus, for instance, in the case of English, the word "blue" tends to co-occur with the word "colour". In this respect, the PPMI^3 attempts to capture the same type of semantic relatedness as the TFIDF metric we described in the previous section.

However, the PPMI^3 matrix allows us to access *second-order co-occurrence* (also called *paradigmatic association*) as well.[33] This means that, after a subsequent transformation and analysis of the matrix, we are able to capture the semantic association between words that perhaps do not occur so often together but do tend to co-occur with similar third-words. Thus, there might be a strong paradigmatic association between the words "blue" and "green", since they both co-occur with a third word "colour" and a number of other colour-related third words. In principle, we can measure this sort of semantic relatedness between any two words by comparing the row vectors corresponding to them within the PPMI^3 matrix. In fact, this sort of vector comparison lies at the core of vector semantics as such, and gives it its name.

However, to make the vector comparison more robust, we further employ Singular Value Decomposition (SVD) to reduce their dimensionality, i.e. we transform them from sparse high-dimensional vectors

[29] Church / Hanks (1990), 22–29.

[30] For all the subsequent analyses, see Kaše / Linka (2021), scripts/5_VECTORS.ipynb (Last access 30.08.2021).

[31] Levy et al. (2015), 211–25.

[32] Role / Nadif (2011), 218–23.

[33] For the difference between first-order and second-order co-occurrence, see Jurafsky / Martin (2020), 274–75 and Schütze / Pedersen (1993), 104–13.

with 2,033 dimensions to lower-dimensional (denser) vectors with 250 dimensions.[34] The outcome is a PPMI^{3}SVD matrix, in which each row corresponds to a 250-dimensional vector representation of a word. Subsequently, we employ cosine similarity to construct a similarity matrix comparing any two-row vectors against one another.[35] Using this similarity matrix, for any word we choose we can easily extract a certain number of the most similar words to it, i.e. its nearest neighbours. As we have already mentioned, these similarities between words attempt to capture the so-called paradigmatic association between them. It has been repeatedly demonstrated that, when trained on large and representative language corpora, this sort of method is able to automatically detect synonymity and some other types of semantic relatedness – a capability that might be evaluated against benchmark tests based on manually coded data.[36]

Results

Corpus overview and document distances

The *Corpus Hippocraticum* (*CH*) extracted from LAGT consists of 52 works.[37] These works are formed by 24,456 sentences and 171,332 lemmatised words tagged either as nouns, proper names, adjectives or verbs.[38] To obtain a basic overview of the corpus, we have produced Figure 1, which plots distances between individual works in *CH* based on similarities and dissimilarities in their vocabulary. The term vocabulary here refers to this subselection of lemmatised words. Works depicted closer to each other tend to share more words than works that are farther from each other.

Upon analysis of Figure 1, we see that it produces some local clusters of works. For instance, quite unsurprisingly, on the left side of the figure we see very close to each other two works which have been classified by Jouanna as “Dietetics”.[39] This suggests that the method performs well in capturing this thematic relatedness. Furthermore, on the right side, we see a cluster of works formed by five texts which have been classified as “Surgical”. Again, drawing on the vocabulary, our measurement properly captures that these five texts are indeed related. At the top of the figure, there is another relatively homogeneous cluster (*Lex*, *De decente habitu*, *De arte*, *Praeceptiones*, *De medico*, *Epistulae*), which, however, does not fall under any single category proposed by Jouanna. Yet, thematically, these writings appear to be related, since all of them somehow concern the profession and social role of the physician.

[34] In the context of vector semantics, SVD was originally popularized by Latent Semantic Analysis, where it serves to reduce the dimensionality of a word-document matrix (see Deerwester et al. [1990], 391–407). Here we employ it to reduce the dimensionality of the PPMI^3 matrix, which might be considered a weighted variant of the word-word co-occurrence matrix. For the same approach and its rationale, see Levy et al. (2015), 211–25.

[35] This analysis shares several features with the analysis of distances we have introduced above. The main difference is that there the vectors represented works, whereas here they represent words.

[36] See e.g. Levy et al. (2015), 211–25; Sahlgren / Lenci (2016); Baroni et al. (2014), 238–47.

[37] 19 works are included in the Perseus Digital Library; the rest originate from the First Thousand Years of Greek project.

[38] In total, the corpus consists of 333,443 raw words. For more details, see Kaše / Linka (2021), scripts/1_EXTRACTING-CORPORA.ipynb (Last access 30.08.2021).

[39] In Figure 1, we adopt a classification proposed by Jouanna (1999), 66–71. He acknowledges that particular writings of the corpus vary in both date and authorship, and that despite it being problematic to categorize the corpus, he attempts to do so and classifies particular writings into groups indicated on the right in Figure 1. This classification is based on the criteria of content and the date of composition. Other authors propose different classifications, emphasising different groups of writing; thus, the discussion about the classification of *CH* is ongoing (see e.g. Craik [2015]), xiv–xxxv). For comparison, we have also generated a figure using Craik’s categories; see Kaše / Linka (2021), figures/c_hip_distances_by_cat_craik.png (Last access 30.08.2021).

It is encouraging to see that our method is able to capture this aspect as well. At the bottom right, we can see another strongly distinguishable group, formed mainly by works classified as “Later” or “Other”. There are also two other works from the category of “Female medicine”. This cluster of works, which represents different categories, can be explained by several factors: It can either mean that at least some of the works classified by Jouanna as “Later” and “Other” are indeed also related to the topic of “Female medicine”, or that the two works from the category of “Female medicine” reveal substantial similarities with some later works. We are not able to decide which is the case here, since it would either require a detailed close reading of the texts or an employment of another CTA (computational text analysis) method, e.g. a stylometric analysis. However, both would divert us from the main topic of this article, which is the understanding of pain.

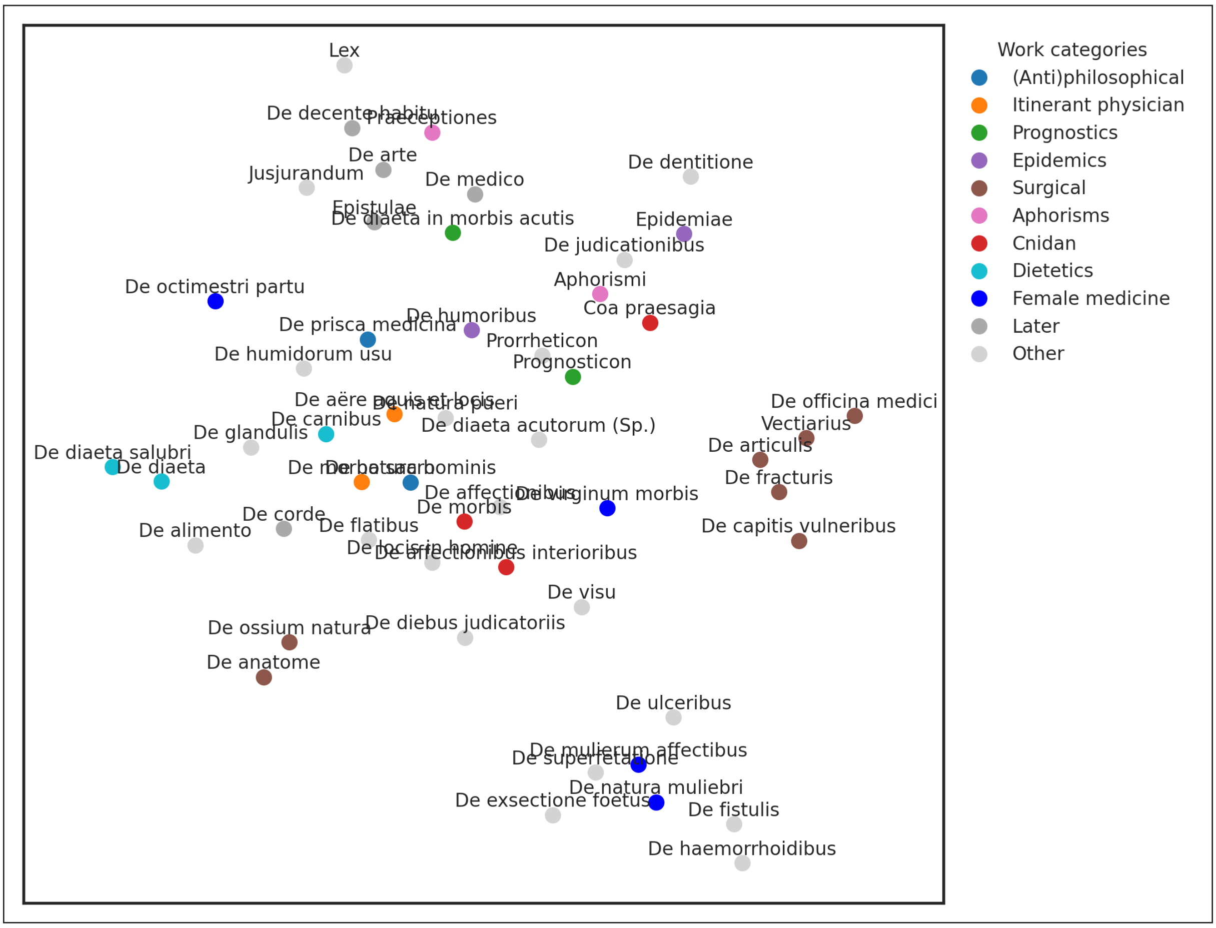


Fig. 1: Work distances based on shared vocabulary.

Taken together, Figure 1 helps us demonstrate several different things. First, it allows us here to validate our overall approach, since the work clustering obtained by this method might easily be evaluated against any work classification offered by experts conducting close reading of the sources. In that respect, we should realise that the classification of works within *CH* is an important prerequisite for almost any inquiry concerned by the history of ancient medicine, given the fact that *CH* is a heterogeneous corpus of works written by different authors over the span of more than a century.[40] At the same time, it can also direct future research, identifying subtle similarities of works that are otherwise treated separately, as in the case of the cluster at the bottom-right of the figure discussed above. However, we should also not ignore the limitations of this particular method. Firstly, it completely ignores word order, employing what is known as a bag-of-words approach.[41] This substantially constrains the possibility of inferring

[40] Craik (2015), xxiv–xxviii.

[41] Jurafsky / Martin (2017), 76.

anything substantial concerning the semantics, since the meaning of words is determined by their context of usage on the level of sentences etc., as captured by the DSM. Secondly, here we focus exclusively on a subselection of lemmatised words, namely nouns, proper names, adjectives and verbs. This naturally flattens any differences in style, which are typically mirrored in the usage of function words like καί, δέ, μέν or τε and which are therefore commonly used in stylometric analyses for authorship attribution.[42]

Pain words across work categories and sentences

After the analysis of work distances, we proceed to the problem of pain in *CH*. In this case, we have to focus on usage of the four pain-words. Our dataset contains 657 instances of πόνο*, 645 instances of ὀδύν*, 315 instances of ἄλγ*, and 58 instances of λύπ*.[43] Thus πόνο* and ὀδύν* appear to be the most frequent ones, while λύπ* tends to be used only rarely. Remarkably, the proportion of usage of these word families is completely different than the one we observe in other ancient Greek texts from a similar period, which represent a different genre. For instance, from the four pain-words, Aristotle most often uses λύπ* (406 instances), followed by πόνο* (103 instances); there are only 34 instances of ἄλγ* and 3 instances of ὀδύν*.[44] Furthermore, as shown in Figure 2, the proportional distribution of the pain-words also broadly varies across individual work categories within the corpus.[45]

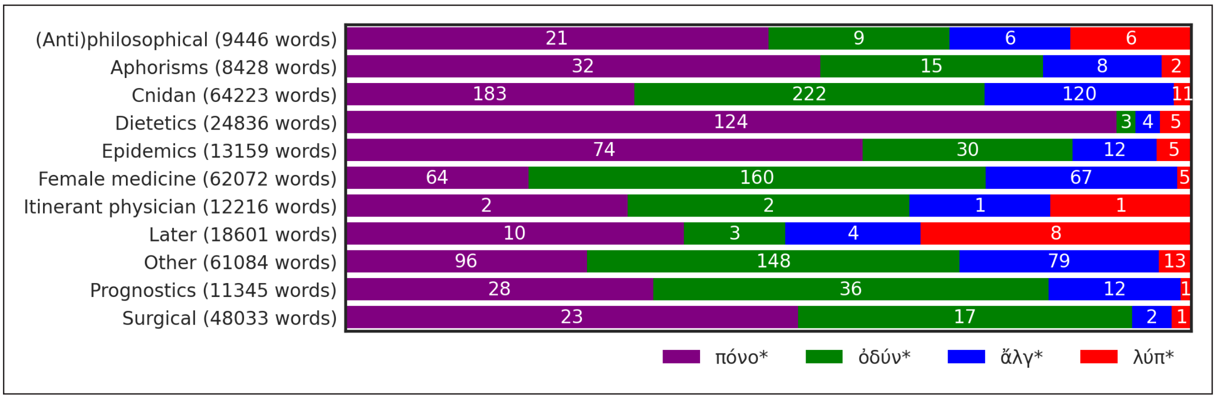


Fig. 2: Ratios of pain-words across work categories by Jouanna.

Figure 3 plots 20 words with the highest TFIDF scores within all sentences containing the individual pain-words. We have manually classified the terms into several categories (pain-word, pathological state, body and its parts, dietetic term, quality, and general term) and differentiated these terms by colours (see legend on Figure 3).[46]

[42] Koentges (2020), 211–41.

[43] In TLG, we can find 891 instances of ὀδύν*, 709 instances of πόνο*, 379 instances of ἄλγ*, and 60 instances of λύπ*. This numerical difference has at least three reasons: Firstly, the TLG search engine includes composite words like κεφαλαλγία, while we focus only on words beginning with the root. Secondly, TLG employs different lemmatisation. Finally, TLG includes editions of some works that are different from ones available via open resources.

[44] For the extraction of pain-words and comparison with Corpus Aristotelicum, see Kaše / Linka (2021), scripts/2_EXPLORATIONS+REPLACEMENTS.ipynb (Last access 31.08.2021).

[45] For the proportion of pain-words across the work categories proposed by Craik, see Kaše / Linka (2021), figures/c_hip_ratios_by_cat_craik.png (Last access 31.08.2021).

[46] For details, see Kaše / Linka (2021), scripts/4_PAIN-SENTENCES.ipynb (Last access 31.08.2021). For the full list of terms classified by categories and accompanied by automatic translations, see Kaše / Linka (2021), data/terms_by_category.csv (Last access 31.08.2021).

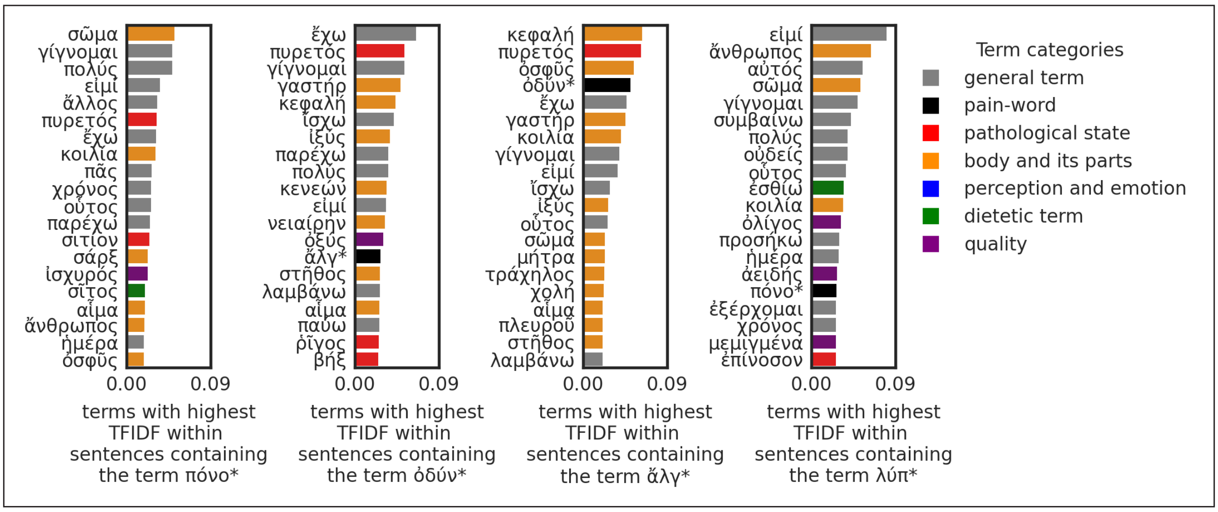


Fig. 3: 20 terms with highest TFIDF within sentences containing the pain-words.

We can see that the terms listed in Figure 3 substantially overlap between the four subplots. However, there are also some remarkable differences. From the four word-families, ἄλγ* seems to co-occur very frequently with individual bodily parts and constituents (12 from 20 terms with the highest TFIDF value), followed by ὀδύν* (7 terms) and πόνο* (6 times).[47] The usage of πόνο* tends to be more general, and is associated with terms like σῶμα ("body"), σάρξ ("flesh") or αἷμα ("blood"). Looking at this data, λύπ* appears to be a term from a slightly different semantic domain, only marginally connected with the somatic and medical domain. This is unsurprising given the fact that λύπ* in classical Greek literature usually denotes sorrow or some other negatively evaluated emotional state.[48] Thus, this analysis of sentences using the TFIDF algorithm gives us some preliminary insights concerning the contexts in which pain-words appear. The advantage of this method is that it is computationally rather straightforward and easy to interpret. However, it does not allow us to go as deep concerning the semantics of the terms under scrutiny. This requires the adoption of more advanced methods, which will be the subject of the following section.

Distributional Semantics and Word Embeddings

By calculating the PPMI^3 value for each possible word pair of all words appearing in at least 5 works within the corpus, we obtained a square matrix of 2,033 rows and 2,033 columns. Subsequently, for each of the pain-words, we used this matrix to extract 20 words having the highest PPMI^3 association with them (see Figure 4). It should not surprise us that the results are highly comparable to the results we obtained using the TFIDF algorithm (see Figure 3). Since both measures attempt to capture the same type of semantic relatedness, we can consider this observation as a sort of validation of this second, more complex measure. It is important, since the PPMI^3 matrix serves us here as a middle step in the construction of a PPMI^{3}SVD matrix, which we can use to calculate word vector similarities in an attempt to capture the paradigmatic association between any two words of our interest.

[47] Remarkably, the method captures this feature even while it does not include words like κεφαλαλγία.

[48] The sense of this word becomes broader in the context of the Greek tragedy, and its authors – Aeschylus, Sophocles, Euripides – also use it in the sense of mental pain. However, it is only in Plato and Aristotle where λύπ* is used for denoting physical pain as well, and it works as a general term for pain in opposition to ἡδονή (pleasure). See Cheng (2019), 47–71. Also, our close reading analysis introduced below shows that in *CH*, λύπ* usually keeps its non-physical-pain sense, even though there are some rare exceptions.

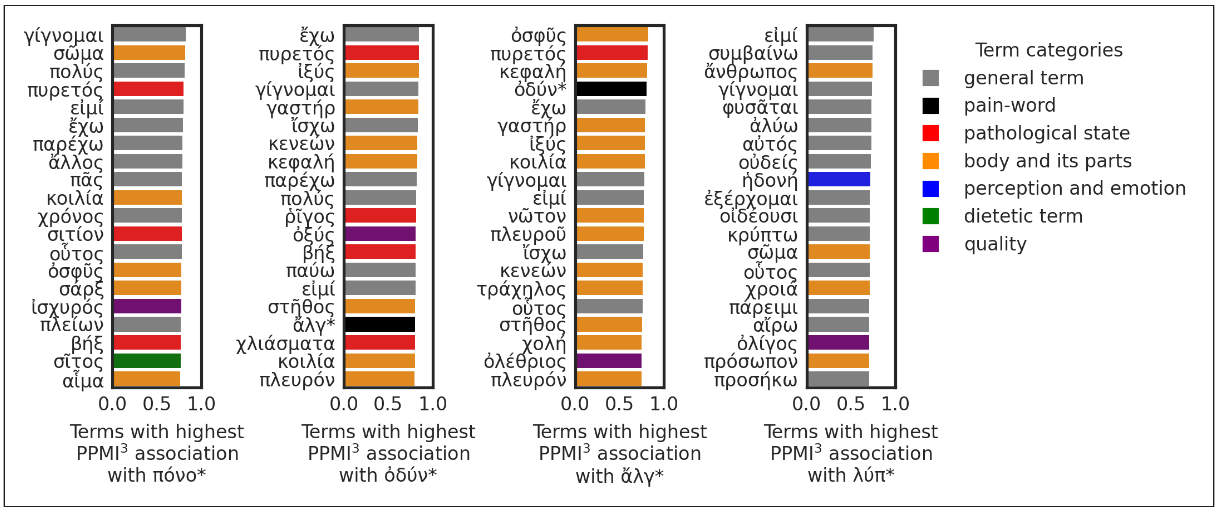


Fig. 4: Pain-words coupled with 20 terms with highest PPMI^3 association with them.

Figure 5 is based on the cosine similarity of words within the PPMI^{3}SVD matrix, in which each row corresponds to a 250-dimensional vector representation of a word. In particular, it contains the 20 nearest neighbours for each of the pain-words together with horizontal bars expressing a cosine similarity score on a scale from 0 to 1. Firstly, when looking at the third column, we see that ἄλγ* is no longer as strongly associated with the body and its parts as was the case of TFIDF and the original PPMI^3 matrix scores (cf. Figure 3 and 4). This should not surprise us, since we are now capturing the second-order (paradigmatic) association and not the first-order (syntagmatic) co-occurrence. Following this, we see in the third subplot that the nearest neighbour of ἄλγ* is ὀδύν*. At the same time, we observe in the second subplot that in the case of ὀδύν*, ἄλγ* occupies the 9th position. The score associating ὀδύν* and ἄλγ* is the same in both cases, but in the case of ὀδύν* there are other terms with higher scores.[49]

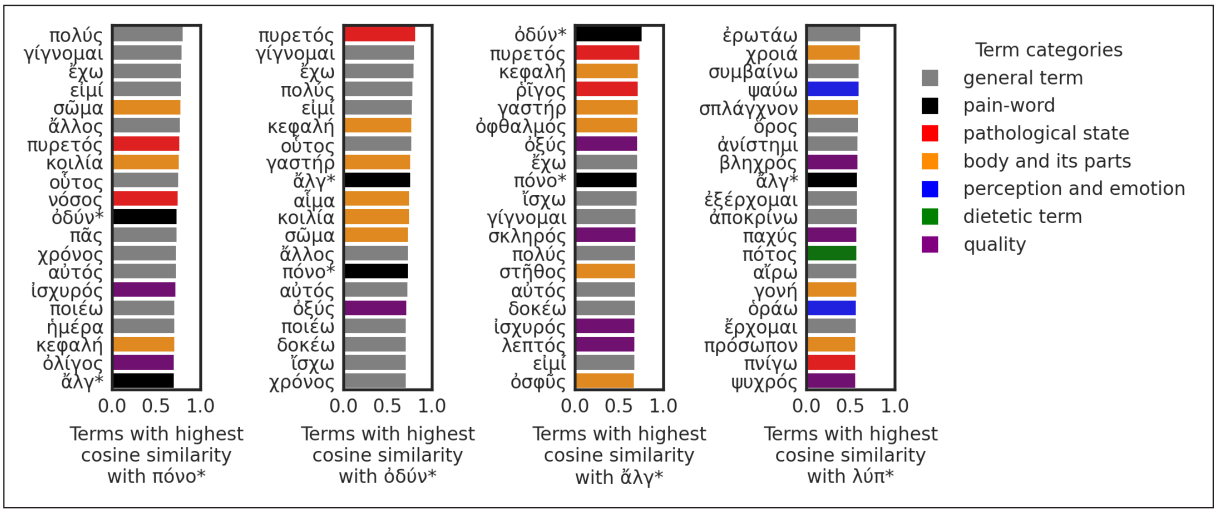


Fig. 5: Pain-words coupled with 20 terms with the highest cosine similarity of vectors based on the PPMI^{3}SVD.

Taken together, there seems to be much overlap between ὀδύν* and ἄλγ*. In both cases, we see a very strong association with πυρετός ("fever"). Both terms reveal a highly medicine-specific context without any clear semantic difference. πόνο* also reveals some association with πυρετός, but the predominance of general terms suggests that the semantic context is slightly different. When we look at the λύπ* column, it appears that we are deviating even farther from the medicine context than in the case of πόνο*.

[49] Again, we observe here a significant number of general terms in the figure. It is a consequence of the PPMI^3 metric, which we used to construct the PPMI^{3}SVD matrix.

Again, as we have already discussed above, this is unsurprising because λύπ* was originally used in the sense of sorrow. The results of Figure 5 will be elaborated upon further within the Discussion.

Relying on the same data that we used for the creation of Figure 5, we can proceed further with another visualisation, which will be to a certain extent similar to the one we used for plotting distances between individual works by drawing on their shared vocabulary. This time we will plot distances between words by inverting similarity scores from the PPMI^{3}SVD similarity matrix. As in the case of work distances, we firstly apply tSNE to project the data from the distance matrix into a 2-dimensional space. Subsequently, we plot these data using a scatter plot, a standard way to visualise word-embeddings. However, since there are 2,033 data points (i.e. words), it is not possible to plot all of them in a meaningful way together within one plot. Therefore, in Figure 6, we instead introduce a series of four subplots.

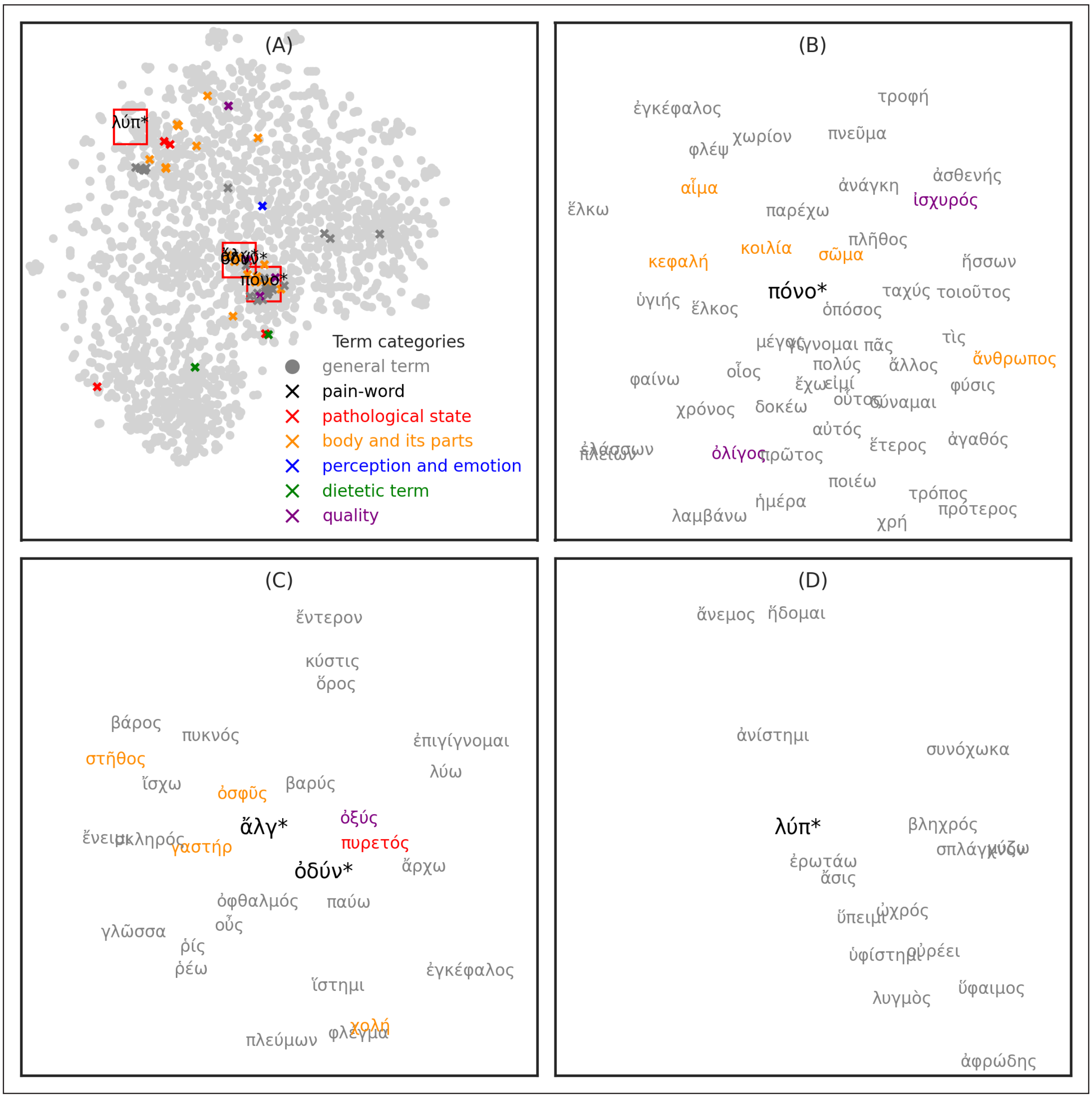


Fig. 6: Word-embeddings based on the PPMI^{3}SVD matrix. Subplots (B), (C) and (D) represent cutouts containing the four pain-words.

Subplot (A) gives us a general overview of the spatial distribution of the 2,033 words within the model. This distribution is based on distances between these words within the PPMI^{3}SVD distance matrix (an inverted version of the PPMI^{3}SVD similarity matrix). Points depicted using a cross sign depict words already contained in Figures 4 and 5. As we move further from the middle of the figure, we can identify some more densely clustered groups of points corresponding to semantically closely related groups of words. We also see that three of the pain-words appear rather close to the centre. This suggests that these words are strongly connected with the rest of the corpus, appearing in more than one specific context. In that respect, λύπ* appears to be much less anchored within it, which also reflects its substantially lower frequency. The subplot (A) further depicts three squares surrounding the pain-words, which are used for a subselection of data for subplots (B), (C), and (D).

In subplot (C), we see that ἄλγ* and ὀδύν* appear very close to each other (this is also seen in the Figure 5). This also allows us to capture their neighborhood using one subplot. Furthermore, we also observe here the close association with πυρετός, which appears to be much stronger than between πόνο* and πυρετός. Finally, we can identify here a number of terms from the category of the body and its parts. The neighbourhood of the term πόνο* as depicted by subplot (B) seems to be preoccupied by semantic general terms which are commonly the most frequent terms in the corpus as a whole. The subplot with λύπ* (D) is extracted from a much less densely populated part of the embedding. This also helps us to understand why its similarity values with its closest neighbours as depicted in the fourth subplot of Figure 5 are comparatively much lower than the values we observe in other subplots. Taken together, it seems that the usage of λύπ* in *CH* does not reveal any specific semantic context, what is also caused by its limited extent of usage.

Discussion

In the previous section, we have captured a significant semantic association between ἄλγ* and ὀδύν*. The similarity of the two pain words to each other is clearly manifested in Figure 6 as well as in Figures 3–5. Both terms are closely associated with bodily organs or pathological states (see especially Figures 3–5, where the connection to bodily organs is substantially stronger than in the case of the other pain words). These insights can be further validated and elaborated by close reading of the texts. When we go through various thematically dissimilar texts, for example *Coa praesagia*, *De fracturis*, *De natura muliebri*, *Prognosticon*, or *Epidemiae*, we find ἄλγ* and ὀδύν* used usually as examples of pain occurring in some particular body part as a result of an illness or other pathological state.[50] It is worth mentioning that ὀδύν* maintains the sense of a specific physical pain even in the treatises which are more theoretical and general, for example *De natura hominis* or *De prisca medicina*,[51] whereas ἄλγ* can be used in these types of treatises as a general term denoting pain, by which the author proposes his theory of the nature of pain.[52]

Whereas in the case of ἄλγ* and ὀδύν* the DSM analysis reveals some clear connections to bodily organs or pathological states, in the case of πόνο* the results are less decisive. We have seen above that this word is closer to general terms rather than to some special medical vocabulary (see especially Figure 6 [B]). Of course, this word is related to other medical terms, too, but remarkably, Figure 3–5 depict it as being closely associated with rather general terms such as σῶμα, rather than to some particular bodily organ. Yet, when we conducted a close reading of some representative selection of Hippocratic texts, we

[50] Hippocr. *Coac*. 18.1, 195.1, 265.5, 274.7; *Fract*. 7.2, 9.21, 17.6; *Nat. Mul*. 2.7, 5.2–4, 5.2, 6.2–3, 7.3, 18.1; *Progn*. 5, 7, 19, 24; *Epid*. 1.2.6.1–14, 1.3.13.17, 1.1.3.26.

[51] Hippocr. *Nat. Hom*. 4.10–14, 11.13, 11.36, 12.2, 15.2, 12.2, 15.2, *Med. Vet*. 19.5, 22.51.

[52] Hippocr. *Nat. Hom*. 2.8–12; 4.3–5.

found πόνο* used in a way very similar to ἄλγ* and ὀδύν*, i.e. in connection with a bodily organ and a pathological state.[53] Thus, we expected that this association will be apparent in the DSM analysis as well. However, we must take notice of the fact that 107 of 657 instances of πόνο* in the whole corpus appear in *De diaeta*, where it has a meaning different than pain. In this dietetic work, πόνο* usually designates exercise.

Thus, to explore the possibility that the overall meaning of this term is substantially influenced by this one work, we re-ran the whole DSM analysis, this time without *De diaeta*.[54] In this version, we found the term πόνο* to be more closely related to the other pain-words, especially ὀδύν*.[55] Thus, it appears that in the case of πόνο*, the overall results are substantially influenced by this particular writing and the specific meaning of πόνο* in it. For instance, we can also see in Figures 3–5 that πόνο* is connected to some temporal attributes such as ἡμέρα or χρόνος, which should not surprise us, because time and duration play an important role in the dietetics. Nevertheless, taken together, it seems that πόνο* has a slightly broader meaning than ἄλγ* and ὀδύν* in *CH*, a feature which is captured within the DSM analysis by its close association with more general terms. This feature is also recognised by at least some translators, who choose to render it as "suffering" or "souffrance".[56]

The DSM analysis of ἄλγ*, ὀδύν* and – to some degree – πόνο* seems to support an interpretation of the problem of pain in *CH* made by some scholars over the past thirty years.[57] They all agree that pain in *CH* figures as a symptom of illness and that it is usually connected with a concrete bodily organ or area. We believe that the close association between pain-words on the one hand and bodily organs and pathological states on the other captured by the DSM analysis supports this claim. It is of interest that all pain-words can relate to various words specifying the quality of pain (sharp, intensive etc.), which, possibly, says something about how the patients classified their pain (this is most noticeable in Figure 5). Yet, with the methods we use in this paper, it is difficult to elaborate on the problem of how the patients felt their pain. To enhance this question, we would need to focus more on semantic and psychological analyses of *CH*.

As we have already mentioned several times, the word λύπ* occupies a specific position within the corpus, which is most clearly seen in Figure 6 (A). Figures 3–5 reveal that λύπ* is not associated as much with bodily organs or pathological states, and is more connected with words of other types, e.g. ἡδονή (see Figure 4). The word λύπ* is also the only pain-word connected to sense-perception (see Figure 5), a trend which is also documented in the philosophical literature of the time.[58] Furthermore, it is noticeable that in the case of Figure 3, both λύπ* and πόνο* maintain a strong connection to general terms like σῶμα ("body"), ἄνθρωπος ("human being"), ἡμέρα ("day") or χρόνος ("time"). Thus, it seems that λύπ* is usually not meant in the sense of a concrete physical pain, an observation which is evidenced by close reading as well. In the majority of writings, this word is either completely absent or used only exceptionally. Even in the works where it is used more often, it usually means an emotion of sorrow[59] or pain or

[53] *Fract*. 2.32, 3.8, 5.30, 6.8. *Coac*. 31.1–3, 76.3, 138.2, 139.2–3; *Nat. Mul*. 5.4, 12.14, 18.2, 23.1. *Epid.* 1.2.6.5–11, 1.2.3.4.123, 1.3.13(2).25, 1.3.13(4).5–7; *Progn*. 5, 11, 19.

[54] See Kaše / Linka (2021), scripts/6_VECTORS_without-de-diaeta.ipynb (Last access 31.08.2021).

[55] The 20 terms with the highest PPMI^3 score were: γίγνομαι, πολύς, σῶμα, πυρετός, εἰμί, ἔχω, ἄλλος, κοιλία, πᾶς, παρέχω, ὀσφῦς, βήξ, χρόνος, ἰσχυρός, οὗτος, αἷμα, ὀδύν*, ὀξύς, κεφαλή, τράχηλος.

[56] See especially theoretical writings like *Vict*., *Nat. Hom*., *Med. Vet*. translated by W. H. S. Jones and E. Littré.

[57] King (1998); Horden (1999), 295–315; Rey (1995).

[58] For Aristotle, for example, sense-perception is a necessary condition for feeling pain and pleasure, and the relation between pain, pleasure and perception is an important topic in Aristotelian scholarship. See *DA* II, 3, 414a32–b16. Cf. Corcilius (2008), 79–82.

[59] Hippocr. *Epid*. 3.3.17(11)3, 3.3.17(15)3.

suffering in general without any explicit connection to a bodily organ or pathological state.[60] However, some moderation is required in ascribing λύπ* to any specific context in *CH*, because of its scarcity and relative distance to other terms (see the interpretation of Figure 6 [B] above).

We should not overlook that our methods are not able to capture some important specifics and exceptions which occur in some particularly important writings of *CH*. Especially in writings such as *Nat. Hom*., *Med. Vet*. and *Vict*., we find intriguing passages about the nature of pain, its generation and further scientific and philosophical implications. However, in *CH* as a whole, the treatises containing an explicated theory of pain make up a minority.[61] In this respect, the value of the DSM analysis lies in its capability to look at the corpus as a whole, without being biased by a few writings that represent an exception. Nevertheless, if we are interested in the problem of pain not in the perspective of the whole corpus, but, for instance, in the treatises particularly influential for the reception of Hippocrates in Western thought, it is important not to overlook some intriguing claims connected to the theory of pain presented in them.[62]

In the future, we envision the possibility of employing other computational text analysis approaches when studying pain in *CH*. In particular, stylometric analysis is a promising research pathway to evaluate some hypotheses discussed within the scholarship. Rey, for example, claims that there is a difference between ὀδύν* and πόνο* based on the prepositions with which these words occur: ὀδύν* usually occurs together with more concrete prepositions, so it is a more precisely localised type of pain, whereas πόνο* denotes a more general type of pain because it is connected with prepositions which are not particularly specific.[63] Without a doubt, it is possible to quantitatively evaluate such claims without employing any advanced computational techniques, e.g. by exploring available word indices and concordances. However, computational stylometry allows us to do this in a more controlled fashion, comparing a large number of features at once. Furthermore, a more complex distributional semantic analysis could also help us evaluate King's claim that πόνο* usually means natural pain (for example birth pain), whereas ὀδύν* unnatural pain (being a result of some damage to the organism).[64] Finally, the methodological framework we employed here can easily be transferred and applied to other comparable or even much larger textual corpora of ancient Greek texts. Thus, for instance, we could analyse the understanding of pain in *Corpus Aristotelicum* or *Corpus Galenicum*, both of which are covered by the LAGT dataset that we have used here. Furthermore, the algorithms in the core of our scripts might also be modified and reused by other scholars to study different topics.

[60] Hippocr. *Med. Vet*. 14.23–28; *Vict*. 15.5–6. Thus, it seems that λύπ* in *CH* has a different meaning than in classical philosophical literature. Only in works like *Med. Vet*. and *Vict*. is the meaning similar. See for example Aristotle, *EN* 1152b1–8; 1153b1–4; 1154a22–31; Plato, *Gorg*. 492a–499a; *Phlb*. 31a–34a, 44a–45a; *Resp*. 583b–584b; *Phaed*. 65b–c, 68e–69b, 83d–84e. For broader discussion about pleasure and pain in Plato and Aristotle, see Cheng (2015); Frede (2016), 255–76; Wolfsdorf (2013). For the semantics of pain in Aristotle and his contemporaries (including *CH*), see Cheng (2019). Plato and Aristotle, however, use λύπ* also in cases where Hippocratic authors would use different pain-words, i.e. in the case of a specific bodily pain.

[61] This is emphasised by Horden (1999), 295–315, who underlines that in respect to pain, *CH* differs from the philosophical corpora of classical antiquity in its absence of any theoretical conception of pain.

[62] For the theory and origins of pain as well as its relation to the nature of the human body, see Hippocr. *Nat. Hom*. 2.8–12; *Med. Vet*. 14.23–28; *Vict*. 66.42–46. For the connection between pain, sense-perception and mind, see *Aph*. II.6, II.46.

[63] Rey (1995), 18–19.

[64] King (1998), 267–286.

Conclusion

In this article, we approached the problem of pain in *Corpus Hippocraticum* by combining a distributional semantic analysis of the corpus with the close reading of selected works. We have especially focused on the semantic similarity between pain-words and other relevant terms. Our interpretation indicates that in the case of ἄλγ*, ὀδύν*, there seems to be a shared close association between pain, bodily organs and pathological states. Thus, as far as we deal with these word families, our findings are in accord with the interpretation advocated by some other scholars, who view pain in *CH* as a symptom of a pathological state located within some part of the body. From the same perspective, the meaning of πόνο* tends to be similar, but slightly more general, revealing substantially weaker association with the medical domain; the word λύπ* stands completely aside. We find it remarkable that, even though the Hippocratic authors offer neither an explicit conception of pain nor its definition, we are able to uncover some general features of its understanding typical for the corpus and to capture some semantic differences between the relevant terms.

Text editions

Littré (1839–1861): É. Littré (ed./transl.), Oeuvres complètes d'Hippocrate, Paris 1839–1861, repr. Amsterdam 1961–1962 & 1973–1991.

References

Altszyler et al. (2016): E. Altszyler / M. Sigman / S. Ribeiro / D. F. Slezak, Comparative Study of LSA vs Word2vec Embeddings in Small Corpora: A Case Study in Dreams Database, arXiv [cs.CL] 5 (2016) http://arxiv.org/abs/1610.01520 (Last access 31.08.2021).

Baroni et al. (2014): M. Baroni / G. Dinu / G. Kruszewski, Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2014, 238–47.

Blackwell / Smith (2019): C. W. Blackwell / N. Smith, The CITE Architecture: a conceptual and practical overview, in: M. Berti (ed.), Digital Classical Philology, Ancient Greek and Latin in the Digital Revolution, Berlin 2019, 73–93.

Cerrato et al. (2020): L. Cerrato et al., PerseusDL/canonical-greekLit 0.0.2711 (Version 0.0.2711), Zenodo, http://doi.org/10.5281/zenodo.4067170 (Last access 10.07.2021).

Cheng (2015): W. Cheng, Pleasure and Pain in Context: Aristotle's Dialogue with his Predecessors and Contemporaries, PhD diss. Humboldt Universität Berlin 2015.

Cheng (2019): W. Cheng, Aristotle's vocabulary of pain, Philologus 163(1) (2019), 47–71.

Church / Hanks (1990): K. W. Church / P. Hanks, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics 16 (1990) 22–29.

Corcilius (2008): K. Corcilius, Streben und Bewegen, Aristoteles' Theorie der animalischen Ortsbewegung, Berlin / New York 2008.

Craik (2015): E. M. Craik, The 'Hippocatic' corpus, London / New York 2015.

Crane (1991): G. R. Crane, Generating and Parsing Classical Greek, Literary and Linguistic Computing 6 (1991), 243–245.

Crane et al. (2020): G. R. Crane / L. Mueller / B. Robertson / A. Babeu / L. Cerrato / T. Koentges / R. Lesage / L. Stylianopoulos / J. Tauber, First1kGreek (Version 1.1.5070), Zenodo, http://doi.org/10.5281/zenodo.4091475 (Last access 10.07.2021).

Deerwester et al. (1990): S. Deerwester / S. T. Dumais / G. W. Furnas / T. K. Landauer / R. Harshman, Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. American Society for Information Science 41 (1990), 391–407.

Frede (2016): D. Frede, Pleasure and Pain in Aristotle's Ethics, in: R. Kraut (ed.), The Blackwell Guide to Aristotle's Nicomachean Ethics, Blackwell publishing 2016, 255–76.

Harris (1954): Z. S. Harris, Distributional Structure, Word & World 10 (1954), 146–62.

Horden (1999): P. Horden, Pain in Hippocratic Medinine, in: J. R. Hinnells et al. (eds.), Religion, Health and Suffering, London 1999, 295–315.

Jänicke et al. (2015): S. Jänicke / G. Franzini / M. F. Cheema / G. Scheuermann, in: Eurographics Conference on Visualization (EuroVis) 2015, 1–21.

Jouanna (1999): J. Jouanna, Hippocrates, Baltimore / London 1999.

Jurafsky / Martin (2020): D. Jurafsky / J. H. Martin, Speech and Language Processing 2020, https://web.stanford.edu/~jurafsky/slp3/ (Last access 31.08.2021).

Kaše (2021): V. Kaše, sdam-au/LAGT v1.0.0 (Version v1.0.0), Zenodo, http://doi.org/10.5281/zenodo.4552601 (Last access 10.07.2021).

Kaše / Linka (2021): V. Kaše / V. Linka, PIA – article supplementary (Version v1.0.2), Zenodo, http://doi.org/10.5281/zenodo.5089410 (Last access 10.07.2021).

King (1998): H. King, Hippocrates' Woman, Reading the Female Body in Ancient Greece, London 1998.

King (1999): H. King, Chronic pain and the creation of narrative, in: J. Porter (ed.), Constructions of the Classical Body, Michigan 1999, 269–286.

Koentges (2020): T. Koentges, Measuring Philosophy in the First Thousand Years of Greek Literature. Digital Classics Online 6,2 (2020), 1–23, https://doi.org/10.11588/dco.2020.2.73197 (Last access 31.08.2021).

Lenci (2018): A. Lenci, Distributional Models of Word Meaning, Annu. Rev. Linguist 4 (2018), 151–71.

Levy et al. (2015): O. Levy / Y. Goldberg / I. Dagan, Improving Distributional Similarity with Lessons Learned from Word Embeddings, Transactions of the Association for Computational Linguistics 3 (2015), 211–25.

López / Ramero (2014): F. López / V. Romero, Mastering Python Regular Expressions, Birmingham 2014.

Mikolov et al. (2013): T. Mikolov / I. Sutskever / K. Chen / G. S. Corrado / J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: C. J. C. Burges et al. (eds.), Advances in Neural Information Processing Systems 26 (2013), 3111–3119.

Moretti (2013): F. Moretti, Distant Reading, London / New York 2013.

Rey (1995): R. Rey, The History of Pain, England 1995.

Role / Nadif (2011): F. Role / M. Nadif, Handling the impact of low frequency events on co-occurrence based measures of word similarity, in: J. Filipe et al. (eds.), Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011), Scitepress 2011, 218–23.

Sahlgren / Lenci (2016): M. Sahlgren / A. Lenci, The Effects of Data Size and Frequency Range on Distributional Semantic Models. arXiv [cs.CL] 27 (2016), http://arxiv.org/abs/1609.08293 (Last access 31.08.2021).

Svärd et al. (2020): S. Svärd / T. Alstola / H. Jauhiainen / A. Sahala / K. Lindén, Fear in Akkadian Texts: New Digital Perspectives on Digital Semantics, in: S.–W. Hsu / J. L. Raduà (eds.), The Expression of Emotions in Ancient Egypt and Mesopotamia, Leiden / Boston 2020, 470–502.

Schütze / Pedersen (1993): H. Schütze / J. Pedersen, A vector model for syntagmatic and paradigmatic relatedness, Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research, Oxford 104–113.

Underwood (2017): T. Underwood, A Genealogy of Distant Reading, Digital Humanities Quarterly 11 (2017), 1–12.

van der Maaten / Hinton (2008): L. van der Maaten / G. Hinton, Visualizing Data Using T-SNE, Journal of Machine Learning Research: JMLR 9 (2008), 2579–2605.

van Rossum / Drake (2009): G. van Rossum / F. L. Drake, Python 3 Reference Manual, Scotts Valley 2009.

Wolfsdorf (2013): D. Wolfsdorf, Pleasure in ancient Greek philosophy, Cambridge 2013.

Author contact information[65]

Vojtěch Linka

Department of Philosophy and Religious Studies
Faculty of Arts
Charles University
nám. Jana Palacha 2
11638, Praha 1
Czech Republic
E-Mail: vojtech.p.linka@gmail.com

Vojtěch Kaše

School of Culture and Society – History
Jens Chr. Skous Vej 5
building 1463, 528
8000, Aarhus C
Denmark

E-Mail: kase@cas.au.dk

&

Department of Philosophy
Faculty of Arts
University of West Bohemia
Sedláčkova 19
30614, Plzeň
Czech Republic
E-Mail: kase@kfi.zcu.cz

[65] The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY 4.0.